Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

The role of entropy and polarity in intermolecular contacts in protein crystals

Marcin Cieślik‡ and Zygmunt S. Derewenda*

Department of Molecular Physiology and Biological Physics and the PSI2 Integrated Center for Structure–Function Innovation, University of Virginia, Charlottesville, Virginia 22908, USA

‡ On leave from the Faculty of Biotechnology, Jagiellonian University, Krakow, Poland.

Correspondence e-mail: zsd4n@virginia.edu

Received 5 November 2008
Accepted 14 March 2009

The integrity and X-ray diffraction quality of protein crystals depend on the three-dimensional order of relatively weak but reproducible intermolecular contacts. Despite their importance, relatively little attention has been paid to the chemical and physical nature of these contacts, which are often regarded as stochastic and thus not different from randomly selected protein surface patches. Here, logistic regression was used to analyze crystal contacts in a database of 821 unambiguously monomeric proteins with structures determined to 2.5 Å resolution or better. It is shown that the propensity of a surface residue for incorporation into a crystal contact is not a linear function of its solvent-accessible surface area and that amino acids with low exposed surfaces, which are typically small and hydrophobic, have been underestimated with respect to their contact-forming potential by earlier area-based calculations. For any given solvent-exposed surface, small and hydrophobic residues are more likely to be involved in crystal contacts than large and charged amino acids. Side-chain entropy is the single physicochemical property that is most negatively correlated with the involvement of amino acids in crystal contacts. It is also shown that crystal contacts with larger buried surfaces containing eight or more amino acids have cores that are depleted of polar amino acids.

1. Introduction

Although crystals constitute an essential prerequisite for X-ray diffraction studies, the nature of the intermolecular interactions responsible for the nucleation and growth of protein crystals has historically attracted limited interest. Consequently, relatively few studies of crystal contacts have been reported in the literature, often as a backdrop for the analysis of biologically functional specific protein–protein interfaces (Janin & Rodier, 1995; Carugo & Argos, 1997; Dasgupta *et al.*, 1997; Janin *et al.*, 2008; Bahadur *et al.*, 2004; Zhu *et al.*, 2006). Although the same physical forces are responsible for the stabilization of physiologically relevant protein interfaces and nonspecific crystal contacts, there are significant differences between them. Functional interfaces have been extensively studied and are relatively well understood in terms of structural features and thermodynamics (Moreira *et al.*, 2007), but crystal contacts are less well understood. What has been well established is that they are individually noticably smaller than functional interfaces, but collectively they engage 8–10 neighboring molecules on average and consequently bury a total of ∼30% of the solvent-accessible protein surface (Janin & Rodier, 1995; Carugo & Argos, 1997). However, larger crystal contacts are significantly more ubiquitous than one would expect from a Gaussian

distribution and they are typically formed by twofold symmetry operations (Janin, 1997).

The physicochemical nature of crystal contacts remains somewhat controversial. It was initially argued that they are effectively stochastic, so that their amino-acid composition is indistinguishable from randomly selected surface patches (Carugo & Argos, 1997; Janin & Rodier, 1995). Even for large crystal contacts, careful analyses of the area-based amino-acid composition revealed only minor deviations from random solvent-exposed surfaces (Bahadur *et al.*, 2004). Assuming that crystal contacts are representative of interactions in solution, their stochastic nature should allow the use of isotropic models in quantitative simulations of protein–protein interactions in solution and in the analysis of phase diagrams. However, such approaches have not been very successful (Liu *et al.*, 2007, 2009; Pellicane *et al.*, 2008). This is because proteins are not spherical and have complex surface topologies which preclude truly random interactions. Thus, recent theoretical studies have focused on the concept of anisotropic or 'patch–patch' interactions (Shiryayev *et al.*, 2006; Cheung *et al.*, 2007; Wentzel & Gunton, 2008). It has been postulated that these interactions may be driven at the microscopic level by hydrophobicity, which confers 'stickiness' to specific surface patches (Pellicane *et al.*, 2008).

There is also experimental evidence that surface patches with specific amino-acid compositions mediate anisotropic interactions during nucleation and protein crystal growth. However, side-chain entropy rather than polarity has been invoked as the primary negative determinant; it was argued that patches depleted of residues with high conformational entropy (*e.g.* Lys, Glu) should form thermodynamically favorable crystal contracts owing to a smaller loss of entropy as the amino acids are packed into the contact interface (Longenecker *et al.*, 2001; Mateja *et al.*, 2002; Derewenda, 2004; Derewenda & Vekilov, 2006). Consequently, it has been suggested that systematic surface mutagenesis replacing selected Lys, Glu and Gln residues with Ala should yield more crystallizable protein variants (Cooper *et al.*, 2007; Goldschmidt *et al.*, 2007). This approach, which is referred to as surface-entropy reduction (SER), has been used successfully in numerous cases to generate X-ray-quality protein crystals (to be reviewed elsewhere) and the mutated patches were almost invariably shown to mediate crystal contacts, which is consistent with the notion of specific and anisotropic, rather than stochastic, interactions.

Given the discrepancies between the previously published analyses of crystal contacts and more recent experimental and computational data, we decided to re-evaluate the physicochemical nature of crystal contacts in a PDB-derived database of nonredundant, unambiguously monomeric proteins. We used single and multivariate logistic regression models to examine the propensity of different residue types for incorporation into crystal contacts. In this way, we evaluated how a set of predictor variables (*e.g.* amino-acid size, polarity, charge *etc.*) affect a dichotomous outcome variable, *i.e.* whether a residue is present in a crystal contact or not. The results reveal that different residue types have different propensities for inclusion in crystal contacts, with side-chain entropy as the principal negative selection variable. Moreover, we show that polarity is the principle negative selection variable for the partitioning of amino acids into contact cores. These results provide theoretical validation and support for the concept of SER as an effective method of enhancing protein crystallizability.

2. Materials and methods

2.1. Monomeric protein structure database

To analyze the nature of contacts in protein crystals, we generated a database of unambiguously monomeric protein structures as a subset of the Protein Data Bank (PDB; Berman *et al.*, 2007), in which all protein–protein interfaces are true crystal contacts and have no functional significance. Only structures with one polypeptide chain in the asymmetric unit were considered. Structures with ambiguous quaternary architecture, based on the annotation by the PiQSi database (Levy, 2007), were removed to avoid any potential bias. Each coordinate set had to satisfy the following additional criteria: (i) no nucleic acid or other large peptide ligands could be present, (ii) the resolution was 2.5 Å or better, (iii) there were no missing backbone atoms and (iv) no ambiguous/duplicate atoms were present. We eliminated several entries with nonstandard space groups, coordinates, residue names *etc.* to avoid processing problems. Only the first set of coordinates was retained for residues with alternate conformations so that all information about static disorder was removed. However, residues with multiple conformers typically occur in fully solvent-accessible loops and do not participate in crystal contacts.

The nonredundancy within the data set was achieved by removing entries clustering together at the 95% amino-acid identity level after the last step in the selection. We used the sequence clustering provided in the PDB and generated using the *CD-HIT* algorithm (Li & Godzik, 2006). A liberal identity cutoff was chosen, since nearly identical proteins can generate different crystal forms. Only one entry with the highest rank (the most representative) was retained for each cluster. A structure was replaced by the next in ranking if it failed the downstream processing pipeline for any reason. The final data set contained 821 unique entries. A total of 51 space groups are represented. The list of PDB entries can be found at http://ginsberg.med.virginia.edu/Files/monomeric_pdb_ids.txt.

2.2. *ZenPDB* in the file-processing pipeline

To carry out the analysis in an automatic mode, we developed *ZenPDB*, a module for the Python programming language. *ZenPDB* integrates a number of open-source software packages, providing a uniform interface for common structural biology tasks. Its possible applications are broader than the application described in this paper and include a robust PDB-file parser and writer, a hierarchical structure class for representing, manipulating and analyzing structural data, interfaces to multiple external libraries and binary

applications, methods and functions for common processing needs, including pipelining capabilities, and a modular design which allows easy addition of modules or features. The source code and documentation are accessible at http://code.google.com/p/zenpdb/.

3. Definitions and concepts

3.1. Accessible surface area

The solvent-accessible surface area (ASA) of a given chemical entity (*i.e.* a protein, residue or atom) is defined by the trace of the center of a spherical probe (r = 1.4 Å) moving around the macromolecule. The ASA values were calculated for each structure within the database using *AREAIMOL* from *CCP*4 (Collaborative Computational Project, Number 4, 1994). Each residue with a calculated solvent ASA > 0 was considered to be part of the protein surface and capable of mediating crystal contacts. The buried solvent-accessible surface area (ΔASA) of a given entity was obtained by subtracting the ASA of that entity (*e.g.* molecule) within the context of the crystal (a complete unit cell together with 26 surrounding unit cells) from the total ASA of the isolated entity.

3.2. Crystal contacts

A combination of two commonly used methods was employed to identify and define crystal contacts for each of the PDB entries. Each crystal contact (and residue/atom part of it) was defined by two properties: the buried ASA and the identity of the neighboring macromolecule (or macromolecules) in the crystal. The buried ASA has a positive value for all residues in contacts. The second property is defined by the symmetry operation and/or lattice translation and is retrieved from the output of the *ACT* program from the *CCP*4 suite. Every residue with buried ASA was identified as part of some crystal contact. If it failed the minimum atom–atom intermolecular distance criterion it was assigned to the same crystal contact as its closest neighbor which forms a contact within 5 Å.

For crystal contacts formed by at least eight residues, the three residues closest to the idealized geometrical center were defined as the contact core. This definition differs from that used in other studies, in which the core is defined as all residues that contain at least one fully buried atom within the contact (Saha *et al.*, 2006). In our opinion, the latter approach has the drawback in that the 'core' of the contact need not even be close to the geometrical center of the contact, might be multipartite and may overestimate the number of residues with long side chains. Our definition identifies residues close to the geometric center, although the cutoff of three amino acids is admittedly arbitrary. All other residues within contacts are defined as the contact rim. Small contacts with fewer than eight residues do not have a core.

Unless otherwise noted, coordinates for contact surfaces and residues used for all Euclidean distance calculations were represented by their idealized geometric centers (centroids). We did not use the C^{α} coordinates because the α-carbon is almost always located at the border of the volume occupied by the residue (Soyer *et al.*, 2000). Fast nearest-neighbor look-up was performed using the ANN library (http://www.cs.umd.edu/~mount/ANN/) and the Python scikits–ANN interface.

Most crystal contacts are binary contacts and by definition involve residues from two symmetry-related molecules (Fig. 1*a*). However, some residues can be buried simultaneously by more than one neighboring molecule and thus belong to two separate binary crystal contacts (Fig. 1*b*). We assign such residues into a separate category of multi-contacts. The fragmentation of crystal contacts was analyzed using both a single-linkage algorithm with a distance threshold of 6.0 Å and an average-linkage algorithm with a 15.0 Å threshold. Both methods gave similar results, although the average-linkage clustering should be more robust with respect to the chaining phenomenon and should yield more spherical contacts (Jain *et al.*, 1999).

3.3. The expected contact area (ECA)

The interactions between different protein molecules in the crystal can be defined by the buried ASA. It is commonly assumed that the probability of a given residue, or atom, being involved in a crystal contact is directly proportional to its ASA

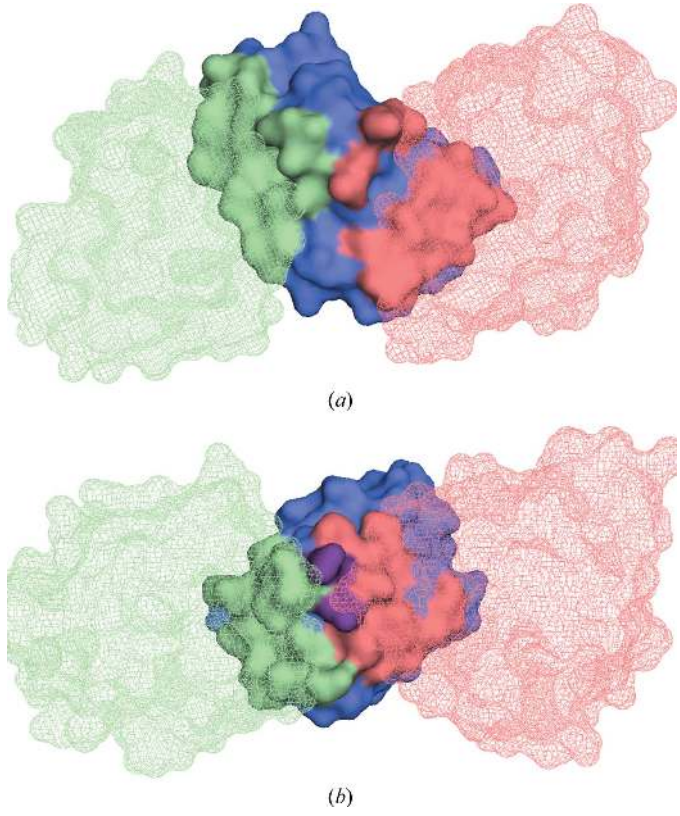

Figure 1
An example of a binary crystal contact (*a*) and a multi-contact (*b*). In each case three neighboring molecules are shown. (*a*) Two binary contacts, of which the red lacks contiguity and the green is contiguous (PDB entry 1be0). The contacts are adjacent but not overlapping. (*b*) Two binary contacts (red and green) overlap to generate a multi-contact (purple) (PDB entry 1k0k). The figure was generated by *PyMOL*.

(Dasgupta *et al.*, 1997; Jelsch *et al.*, 1998). This assumption does not differentiate between equally sized exposed surfaces and partly buried crevices, even though the crystal contact-forming propensities of the two may be substantially different. The fact that large polar residues tend to be highly exposed, whereas hydrophobic residues are abundant in the core of the folded globular protein makes it necessary to establish a relationship between the ASA and the propensity for contact formation and to assess the validity of area-based statistics. To achieve this, the contacts and surface exposures must be calculated for multiple proteins and crystals. We here introduce the concept of the expected contact area (ECA), either of a residue (rECA) or atom (aECA), in a given protein structure. ECA is equivalent to a normalized ASA, *i.e.* independent of packing density. To obtain rECA or aECA values in a given crystal form, the calculated ASA of the residue (or atom) is weighted by the ratio of buried to total ASA of the entire molecule. Thus, the ECA value for any given residue (or atom) is proportional to both its ASA and to the fraction of the surface of the molecule buried by crystal contacts,

$$\mathrm{rECA} = \mathrm{rASA}\left(\frac{\Delta\mathrm{ASA}}{\mathrm{ASA}}\right). \quad (1)$$

3.4. Amino-acid polarity, charge and side-chain entropy

The polarity scale for amino acids used in this study is based on average ranking of amino acids in 38 published hydrophobicity scales (Trinquier & Sanejouand, 1998). This scale assigns low rank to hydrophobic residues (*e.g.* Ile, rank 1) and high rank to hydrophilic residues (*e.g.* Lys, rank 20). Glycine, alanine, tyrosine, histidine and serine are in the middle of the ranking, which reflects their amphiphilic nature. We used the side-chain entropy (SCE) mean scale as defined elsewhere (Doig & Sternberg, 1995), except for the inversion of the sign to obtain a scale that ranks amino acids in an increasing order of side-chain entropy. For logistic regression purposes, charge was considered to be a binary property (*i.e.* charge or no charge) with no sign. Only Asp, Glu, Arg and Lys were considered to be charged. This, of course, is a simplification, as the charged amino acids are correlated with high entropy as well as polarity, and we will discuss this later.

All three scales are shown explicitly in Table 3.

4. Logistic regression

Logistic regression, also known as the maximum-entropy classifier, is used to model a binary (dichotomous) variable as a linear combination of multiple explanatory variables, either numerical or categorical, which can be considered simultaneously. A plot of the probability as a function of some explanatory variable is often nonlinear and S-shaped. The logistic (logit) function is a link function used to transform this S-shaped probability curve into a straight line not confined to the range 0–1. The transformed function can be modeled using a multivariate linear model, but the parameters cannot be estimated using linear regression. Instead, maximum-likelihood estimation is applied. This method involves finding parameter values which give rise to the maximum probability assuming a binomial distribution. The observed frequency (the ratio of outcomes for a range of the explanatory variable) is mathematically the maximum-likelihood estimator.

We use logistic regression to evaluate how physicochemical properties, *i.e.* explanatory variables, of different residues (or sometimes atoms) affect the probability of their being or not being (binary variable) part of a crystal contact. If P_i is the probability that the ith amino acid is a part of a crystal contact, the applied logistic regression models are nested within the following general model,

$$\ln\left[\frac{P_i}{1-P_i}\right] = \alpha + \beta_{\mathrm{polarity}}\mathrm{POLARITY}_i + \beta_{\mathrm{SCE}}\mathrm{SCE}_i + \beta_{\mathrm{charge}}\mathrm{CHARGE}_i + \beta_{\mathrm{rECA}}\mathrm{rECA}_i, \quad (2)$$

where α is the intercept and $\beta_{\mathrm{polarity}}$, β_{SCE} and β_{charge} are the slopes for the explanatory variables, *i.e.* polarity ($\mathrm{POLARITY}_i$), side-chain entropy (SCE_i) and charge (CHARGE_i), respectively.

The general forward selection procedure is applied, in which variables are added to the model starting with the most significant. An analogous general model was used to study the influence of the same set of parameters on the distribution of residues within the crystal contact, where P_i was the probability of the ith residue being located at the core of a crystal contact and all other parameters were exactly as above.

To assess the influence of residue/atom exposure (*i.e.* ASA) on atom contact probability, we used the following model,

$$\ln\left[\frac{P_i}{1-P_i}\right] = \alpha + \beta_{\mathrm{aECA}}\mathrm{aECA}_i + \beta_{\mathrm{rECA}}\mathrm{rECA}, \quad (3)$$

where aECA_i and rECA_i are the atom and residue ECA values for the ith atom accordingly. β_{aECA} and β_{rECA} are the corresponding slopes and α is the intercept.

The slope parameters (β_j for the jth explanatory variable) can be interpreted as the cumulative effect of the jth variable on the log odds ratio of the outcome of the binary variable. It is important to realise that because the units of the explanatory variables are distinctly different, the slope parameters are not normalized and they are not directly comparable. The significance of these parameters in multivariate models was calculated using the Wald test (Z test). For each parameter, the Wald statistic is the square of the ratio of the parameter's value to its standard deviation. The Wald statistics are approximately χ^2 distributed. Improvement between models has been evaluated using the differences in their log likelihoods, assuming a χ^2 distribution (LRT test), *i.e.* p-values. Thus, the lower the p-value associated with a given parameter, the more significant the correlation with the outcome variable. The null model in each comparison was a reduced model with solvent accessibility (rECA or aECA) alone. Parameter estimation and statistical calculations were performed in the R language (http://www.r-project.org) using the Rpy interface (http://rpy.sourceforge.net) for Python.

5. Results

5.1. The size of crystal contact interfaces

First, we analyzed the number and size of crystal contact interfaces in our database. Table 1 shows selected properties, *i.e.* the number of residues involved, overall size and spatial contiguity, binned by the number of residues in a contact. Very small interfaces with up to four residues constitute nearly half of all crystal contacts in our database, with an average buried ASA of less than 100 Å^2. While generally not fragmented, they are predominantly multi-contacts involving more than two protein molecules resulting from overlapping binary contacts. Dense crystal packing is usually associated with a higher percentage of multi-contacts (not shown), often made up of no more than a single residue owing to the geometric restrictions of such contacts. Larger interfaces involving five or more amino acids constitute the remainder and as they increase in size (*i.e.* buried ASA) they become increasingly fragmented and exclusively binary.

5.2. Solvent exposure *versus* contact propensity

Assuming that the probability of a residue being incorporated in a crystal contact is a direct (*i.e.* linear) function of its solvent exposure, the frequencies of amino acids at crystal contacts should be directly proportional to their ASA or, more specifically, to their rECA (see above). To assess whether this is true, we plotted the frequency of residues in crystal contacts as function of their rECA (Fig. 2). Although the frequency increases monotonically with rECA, the dependence is not linear. Most solvent-exposed residues have low rECA values, *i.e.* <15 Å^2, and rarely occur in contacts. As rECA increases, the dependence becomes asymptotic and for residues with the highest rECA values the probability of their participation in a crystal contact exceeds 90%. Less than 1% of all solvent-exposed residues fall into the category with rECA greater than 55 Å^2 (Fig. 2).

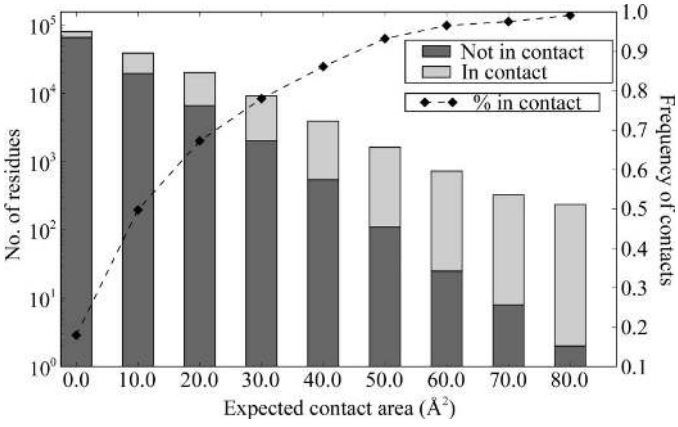

Figure 2
A histogram of the number of residues binned by expected contact area (rECA). For each bin (which is 10 Å^2 wide and is shown centered on the lower limit value) we show both residues in contacts and those not in contacts (note the logarithmic scale for the histogram). The fraction of residues in contact as a function of rECA is shown separately with a linear scale on the right.

Table 1
Topological properties of crystal contacts.

No. of residues†	No. of contacts‡	⟨ΔASA⟩§ (Å^2)	% Split (single linkage)¶	% Split (average linkage)††	% Multi-contacts‡‡
1	2190	60.74	0.0	0.0	80.7
2–4	2590	110.35	13.3	21.2	49.1
5–7	1337	192.30	20.8	13.1	2.9
8–10	1208	283.20	22.9	29.1	0.0
11+	2043	498.39	30.7	76.1	0.0
Total	9368	217.36	16.3	22.8	32.9

† Total number of residues in a contact. ‡ Total number of contacts with a given number of residues in the database. § Mean buried accessible surface area for a given set of contacts. ¶ Percentage of contacts that are split when using the single-linkage algorithm. †† Percentage of contacts that are split when using the average-linkage algorithm. ‡‡ Percentage of contacts that are classified as multi-contact.

The asymptotic dependence of the frequency of residues in crystal contacts on rECA can be rationalized in terms of surface topology, because less exposed residues are more likely to be located in surface crevices and therefore cannot be involved in intermolecular interactions. Thus, the frequencies of individual atoms in crystal contacts should be positively correlated with the degree to which the residue of which the atom is a part is exposed to solvent. To test this assertion, we used a logistic regression model in which the probability of any atom being involved in a crystal contact is a function not only of its expected contact area (aECA), but also of the ECA of the residue to which it belongs (3). In this way, atoms of partly buried residues are distinguished from atoms in solvent-accessible residues. The control (null) model includes only aECA as an explanatory variable. The maximum-likelihood estimates of the α and β parameters of both models are shown in Table 2. We note that the addition of the rECA variable dramatically improves the model, as judged by the p-value, while the relatively high slope, $\beta_{\rm rECA}$, confirms the positive correlation.

5.3. Propensities of amino-acid types for crystal contacts

Next, we asked whether the distribution of amino-acid types as a function of rECA is random. To simplify the analysis, we binned 20 amino acids into five categories, *i.e.* aromatic (Trp, Phe, Tyr, His), aliphatic (Val, Ile, Leu), small (Ala, Ser, Thr, Gly, Ser), charged (Lys, Glu, Asp, Arg) and other (Met, Pro, Asn, Gln), and plotted their observed frequencies in crystal contacts as a function of rECA (Fig. 3*a*). Not surprisingly, small and aliphatic amino acids dominate at the low end of the rECA values, while charged residues dominate at medium and high values. A likely explanation is that all charged residues (Arg, Asp, Glu and Asp) are large and predominantly located on the surface; the inverse argument holds true for small and aliphatic amino acids, which are often partly buried and so their average rECA is relatively smaller.

We then plotted analogous frequencies of residues in crystal contacts (again separately for the five categories; Fig. 3*b*). In each case, we see the same asymptotic dependence on rECA, but for a given value of rECA the charged residues have the lowest probability of occurring within a crystal contact, while

Table 2
Estimates of maximum-likelihood parameters and likelihood ratio test for logistic regression models of crystal contact.

Model	Maximum-likelihood estimates of parameters			
	α†	β_{aECA}‡	β_{rECA}§	p-value¶
aECA	−1.646	0.217	—	
CI 99%††	(−1.657, −1.636)	(0.215, 0.219)	—	
aECA + rECA	−2.157	0.163	0.0360	0.0‡‡
CI 99%	(−2.142, −2.171)	(0.160, 0.165)	(0.0354, 0.0366)	

† Intercept. ‡ Slope of atom expected contact area. § Slope of residue expected contact area. ¶ Probability of log-likelihood ratio test (LRT) for the tested model. †† Estimated 99% confidence interval. ‡‡ The p-value for this model was below the computational threshold and was rounded to 0.0.

the small residues have the highest probability. To obtain a quantitative contact-propensity scale for all 20 amino acids, we evaluated logistic regression models for each amino acid separately, with rECA as the only explanatory variable (1). The resulting α and β parameters for every residue type are shown in Table 3. The β values quantify the contact-forming propensities of individual amino acids and generate a propensity scale. Small and aliphatic residues show the highest propensities for involvement in crystal interfaces and for isosteric or nearly isosteric pairs (*e.g.* Cys and Ser) the more hydrophobic amino acid has a higher contact propensity. In contrast, charged residues (Lys, Glu, Arg, Asp) and two large polar residues (Asn, Gln) show the lowest propensity.

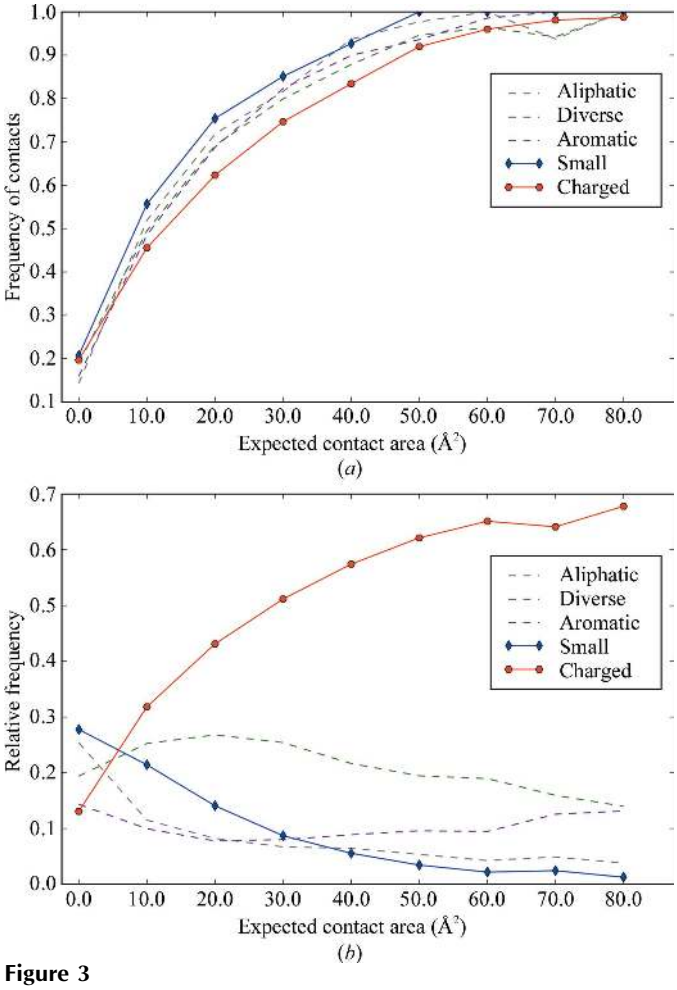

Figure 3
(*a*) Relative frequencies of five categories of amino acids, *i.e.* aliphatic (Val, Leu, Ile), aromatic (Trp, Phe, Tyr, His), small (Ala, Gly, Ser, Thr, Cys), charged (Lys, Arg, Glu, Asp) and other (Asn, Gln, Met, Pro), binned as a function of rECA. The relative frequency in each bin is the ratio of the number of residues of a given type to the total number of residues. (*b*) The fraction of residues involved in crystal contacts as a function of rECA plotted for the five categories as defined above.

5.4. The impact of side-chain entropy, polarity and charge on crystal contact propensities

Given our results showing that different amino acids vary with respect to their propensities for inclusion in crystal contacts, we investigated whether specific physicochemical properties such as polarity, side-chain entropy and charge can account for the observed differences. Again, we used logistic regression models, as defined by (2), to assess the degree to which the inclusion of these explanatory variables improves the expanded models over the simple (null) rECA model.

Table 4(*a*) summarizes the results. The null model assumes that the contact frequencies can be predicted on the basis of the rECA only. Each of the remaining models results from the addition of one of the three additional explanatory variables, *i.e.* polarity, SCE and charge. As judged by the p-values, all three models are significantly better than the null model and thus they are either all individually significant or they are strongly correlated. The β_{rECA} parameter is nearly identical and positive in all models, consistent with positive correlation with rECA. All β parameters for the added variables are negative, consistent with negative correlation of the probability of a residue being involved in a crystal contact with polarity, SCE or the presence of charge. The model which incorporates SCE fits the observed data best, as judged by the p-value. The second best model is that which incorporates charge. However, since most large side chains with high conformational entropy are also charged (Lys, Arg, Glu), charge and side-chain entropy are strongly correlated and consequently both models can be expected to result in similar improvements over the null model. Interestingly, polarity alone has a significantly smaller impact as judged by the p-value, while the β-parameter for polarity is relatively small compared with those estimated for SCE and charge, which is consistent with the low relative importance of this variable. Given that three of the four charged amino acids are also at the top of the polarity scale, one would expect a high correlation between the two variables. However, this is not evident from our analysis. We conclude that SCE is the dominant property, while charge shows statistical significance owing to its strong correlation with SCE.

To obtain a more accurate assessment of the relative importance of each of the three explanatory variables, we constructed a multivariate model which simultaneously incorporates all of them (Table 4*b*). The results are consistent with the individual models and confirm the dominant role of SCE.

Table 3
Estimates of maximum-likelihood parameters for logistic regression models of crystal contacts and a comparison of frequencies of amino acids in crystal contacts (core, rim and total) and total protein surface based on buried ASA.

The side-chain entropy (SCE) and polarity (POL) scales used in the logistic regression models are also shown.

Amino acid	Maximum-likelihood estimates of parameters		Contact core (%)	Contact rim (%)	Contact total (%)	Contact surface (%)	POL§	SCE¶
	β_{rECA}†	α‡						
Gly	0.161 ± 0.010	−1.937 ± 0.098	4.40	4.58	4.55	4.52	11	0.00
Leu	0.151 ± 0.009	−2.324 ± 0.099	6.71	4.11	4.53	4.42	3	0.71
Ile	0.150 ± 0.011	−2.322 ± 0.125	4.12	2.26	2.56	2.41	1	0.76
Val	0.149 ± 0.010	−2.225 ± 0.110	4.72	2.95	3.23	3.18	4	0.43
Ala	0.149 ± 0.009	−2.067 ± 0.101	4.60	4.57	4.58	4.49	9	0.00
Phe	0.146 ± 0.012	−2.256 ± 0.136	4.55	2.43	2.77	2.35	2	0.62
Cys	0.144 ± 0.021	−1.941 ± 0.196	0.81	0.54	0.59	0.66	7	0.85
Tyr	0.130 ± 0.010	−2.098 ± 0.135	5.75	3.58	3.92	3.47	8	1.13
Ser	0.129 ± 0.008	−1.862 ± 0.106	5.51	5.53	5.53	5.52	14	1.11
Met	0.125 ± 0.014	−2.304 ± 0.200	1.97	1.58	1.72	1.51	5	1.46
Trp	0.123 ± 0.016	−2.170 ± 0.211	2.16	1.46	1.58	1.33	6	0.99
Pro	0.118 ± 0.008	−1.870 ± 0.125	5.24	5.49	5.45	5.22	13	0.06
Thr	0.115 ± 0.008	−1.819 ± 0.108	6.35	5.26	5.44	5.50	12	1.08
His	0.105 ± 0.011	−1.900 ± 0.169	2.95	2.46	2.53	2.55	10	0.95
Asn	0.105 ± 0.007	−1.768 ± 0.123	5.57	6.82	6.62	6.31	16	1.03
Asp	0.098 ± 0.006	−1.673 ± 0.107	6.18	8.07	7.77	8.35	19	0.78
Gln	0.094 ± 0.007	−1.711 ± 0.141	5.70	6.13	6.06	5.79	17	1.73
Arg	0.086 ± 0.006	−1.684 ± 0.128	8.60	9.66	9.49	8.61	15	1.88
Glu	0.084 ± 0.005	−1.624 ± 0.112	7.19	10.79	10.21	10.83	18	1.46
Lys	0.074 ± 0.005	−1.545 ± 0.116	6.82	11.64	10.87	13.00	20	1.89

† Slope of residue expected contact area (rECA). ‡ Intercept. § Polarity scale from Trinquier & Sanejouand (1998). ¶ Side-chain entropy scale from Doig & Sternberg (1995).

5.5. The composition of contact cores: the dominant role of polarity

We next asked whether the distribution of amino-acid types within crystal contacts is random or if they are distinctly partitioned between the core and the rim. Fig. 4 shows the distribution of different residue types within crystal contacts, again binned into five categories, as a function of the fraction of rECA buried by those contacts. Assuming that residues with larger buried rECA are located more towards the center of contacts, this is approximately equivalent to the radial distribution of different amino acids within contacts. There are relatively few residues with more than 70% buried rECA, which is a reflection of the average small size of crystal contacts. It is also evident that charged amino acids occur preferentially at the peripheries of crystal contacts and their frequencies in crystal contacts decrease rapidly with increasing rECA. We then analyzed the distribution of the 20 amino acids, as defined by the total buried ASA, in the core and rim of crystal contacts containing eight residues or more, as described in §2 (Table 3*b*). The results are consistent with the radial distribution analysis: the crystal cores are enriched in hydrophobic small amino acids, whereas the rim is enriched in polar and charged residues. To evaluate the significance of polarity, charge and SCE in the partitioning of amino acids between the core and rim, we generated logistic regression models, based on (2), analogous to those calculated for whole crystal contacts (Table 5*a*). The null model evaluates the core contact propensity as a function of rECA only. The β_{rECA} parameter is consistent with a negative correlation between rECA and the residue's probability of being part of the contact core. This implies that on average less exposed residues, irrespective of their actual size, are located at the cores of the contacts. Of the three explanatory variables, polarity shows the most significant negative correlation with the propensity of the residue to partition to the core of a contact. Interestingly, SCE does not appear to be a factor, as the *p*-value for the corresponding model is not significant. This is probably because even peripheral association with a contact may lead to a significant loss of entropy, so that large side chains are disfavored in general both within the rim and at the core.

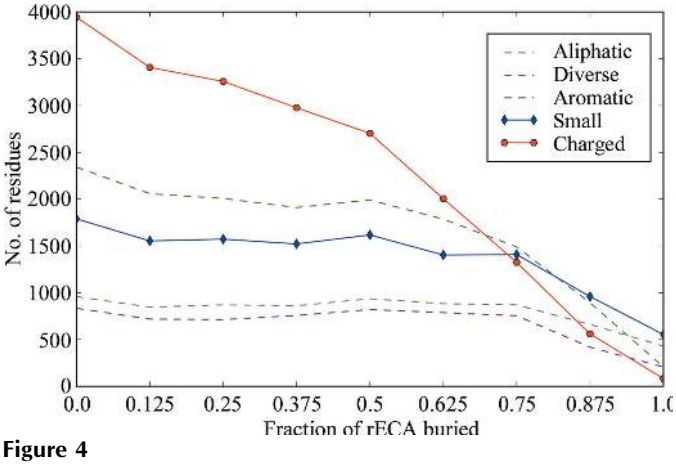

Figure 4
Number of residues in five categories (as defined in Fig. 3) as a function of the fraction of rECA buried in the crystal contacts.

To further validate the results, we calculated a multivariate model which tested the relative significance of all three parameters: polarity, SCE and charge (Table 5*b*). Polarity is by far the most statistically significant parameter negatively correlated with residue partitioning to the contact core. We conclude that such partitioning is driven primarily by hydrophobicity.

6. Discussion

We conducted a statistical analysis of intermolecular contacts in the crystal structures of 821 unambiguously monomeric proteins determined to a resolution of 2.5 Å or better. Our primary objective was to evaluate whether protein crystal

Table 4
Parameters governing the presence of amino acids at crystal contacts.

(*a*) Null and single added variable models.

Model	α†	$\beta_{\rm rECA}$‡	$\beta_{\rm added}$§	p-value¶
rECA	−1.841	0.106	—	
CI 99%††	(−1.867, −1.815)	(0.105, 0.108)	—	
rECA + polarity	−1.789	0.108	−0.00583	1.09×10^{-7}
CI 99%	(−1.825, −1.753)	(0.106, 0.110)	(−0.00867, −0.00300)	
rECA + SCE	−1.665	0.112	−0.265	1.88×10^{-138}
CI 99%	(−1.697, −1.634)	(0.110, 0.113)	(−0.292, −0.237)	
rECA + charge	−1.819	0.111	−0.286	1.45×10^{-86}
CI 99%	(−1.845, −1.793)	(0.109, 0.113)	(−0.324, −0.248)	

(*b*) Multivariate model.

	α†	$\beta_{\rm rECA}$‡	$\beta_{\rm polarity}$ ‡‡	$\beta_{\rm entropy}$ ‡‡	$\beta_{\rm charge}$ ‡‡
	−1.824	0.111	0.018 ($p = 7.1 \times 10^{-35}$ §§)	−0.237 ($p = 1.1 \times 10^{-80}$)	−0.267 ($p = 3.5 \times 10^{-40}$)
CI 99%††	(−1.868, −1.781)	(0.109, 0.113)	(0.014, 0.022)	(−0.270, −0.205)	(−0.319, −0.215)

† Intercept. ‡ Slope of residue expected contact area. § Slope of the added exploratory variable. ¶ Probability of log-likelihood ratio test (LRT) for the tested model. †† Estimated 99% confidence interval. ‡‡ Slope parameter for the respective explanatory variable. §§ Significance of a given parameter obtained with Wald statictics (*Z*-test).

Table 5
Parameters governing the presence of amino acids at the core of crystal contacts.

(*a*) Null model and single added variable models. NS, not significant.

Model	α†	$\beta_{\rm rECA}$‡	$\beta_{\rm added}$§	p-value¶
rECA	−0.694	−0.0540	—	
CI 99%††	(−0.746, −0.642)	(−0.0568, −0.0511)	—	
rECA + polarity	−0.340	−0.0484	−0.0375	4.71×10^{-74}
CI 99%	(−0.411, −0.269)	(−0.0513, −0.0455)	(−0.0428, −0.0322)	
rECA + SCE	−0.717	−0.0545	0.0347	0.0736 (NS)
CI 99%	(−0.779, −0.655)	(−0.0575,−0.0516)	(−0.0153, 0.0847)	
rECA + charge	−0.675	−0.0514	−0.226	5.31×10^{-17}
CI 99%	(−0.727, −0.622)	(−0.0543, −0.0485)	(−0.296, −0.155)	

(*b*) Multivariate model. NS, not significant.

	α†	$\beta_{\rm polarity}$ ‡‡	$\beta_{\rm entropy}$ ‡‡	$\beta_{\rm charge}$ ‡‡
	−0.956 ($p = 2.8 \times 10^{-231}$ §§)	−0.059 ($p = 5.8 \times 10^{-114}$)	NS (0.058) ($p = 1.2 \times 10^{-2}$)	−0.144 ($p = 1.2 \times 10^{-4}$)
CI 99%††	(−1.031, −0.880)	(−0.066, −0.052)	(−0.002, 0.118)	(−0.240, −0.048)

† Intercept. ‡ Slope of residue expected contact area. § Slope of the added exploratory variable. ¶ Probability of log-likelihood ratio test (LRT) for the tested model. †† Estimated 99% confidence interval. ‡‡ Slope parameter for the respective explanatory variable. §§ Significance of a given parameter obtained with Wald statictics (*Z*-test).

contacts are isotropic (*i.e.* stochastic) in nature, as concluded by earlier studies (Janin & Rodier, 1995; Carugo & Argos, 1997), or anisotropic, as suggested by more recent computational and experimental studies (Derewenda & Vekilov, 2006; Pellicane *et al.*, 2008). Isotropic, or random, interactions imply that the surfaces involved are not distinguishable from randomly selected solvent-exposed surfaces of the protein in terms of amino-acid composition. Assuming that the probability of an amino acid being involved in an interaction is directly proportional to its ASA, one can derive a contact-propensity scale using the logarithm of the ratio of the amino acid's area-based frequency in the interface (defined as the buried ASA) to its area-based frequency in the total molecular ASA (Bahadur *et al.*, 2004). Such an area-based composition scale was used to show that large crystal contacts, *i.e.* those generated by twofold rotational symmetry, are slightly enriched in aliphatic and aromatic residues (LIVMFYW) and somewhat depleted of lysine and acidic residues (KED) (Bahadur *et al.*, 2004), although the analysis was not extended to all crystal contacts. However, the area-based approach is based on the premise that the solvent-exposed protein surface, *i.e.* the surface accessible to a water probe, is equivalent to the surface capable of making contact with another molecule. This is not completely true because proteins are characterized by irregular surface landscapes which consist on average of 36% knobs (or protrusions) and 62% clefts (or crevices), with the knobs containing residues that are 30% more likely to be involved in contacts than those in clefts (Albou *et al.*, 2008). Our results also clearly show (Fig. 2) that the contact frequency for residues with small accessible surface area is lower per Å^2 than for more exposed residues.

Notwithstanding, the area-based calculations would still be correct if the amino-acid compositions of knobs and clefts were the same. However, we show explicitly that this is not the case: amino acids with small average ASA (*i.e.* those in clefts) are predominantly hydrophobic and small, while those that are exposed (*i.e.* within knobs) are charged large residues (Fig. 3*a*). Similar results have recently been reported using alpha shapes representations (Albou *et al.*, 2008).

To conclude, the area-based approach underestimates the propensity of small hydrophobic residues for inclusion in crystal contacts. For any given range of ASA, smaller and hydrophobic amino acids actually have a higher relative propensity for involvement in crystal contacts than large charged residues.

In principle, the discrepancy between the solvent-accessible surface and the contact-capable surface can be resolved in area-based calculations by the use of a more stringent lower ASA cutoff value of as high as 30% (Negi & Braun, 2007). In

this way, many of the residues/atoms in clefts are excluded from the protein surface. Such an approach results in the selection of an effective contact-capable surface which is more hydrophilic than the water-accessible surface, but it also generates ambiguity when residues that are not classified as exposed are actually physically incorporated into contacts.

It should be noted that the area-based approach is still applicable if either the selection pressure is very strong (as is the case in evolved biological interfaces) or if frequencies are compared between states which have been selected using the same criteria (such as a comparison of the buried surface area in biological and crystal interfaces). Given that stable biological interfaces developed in response to evolutionary pressure, while crystal contacts are formed by surfaces with no functional significance, the magnitudes of selection are dramatically different. Consequently, area-based composition analysis allows distinction between them and the observed quantitative differences are valid (Zhu *et al.*, 2006). However, this methodology is not sensitive enough for comparisons of crystal contacts and random surface patches, where the differences are more subtle.

Here, we propose an alternative approach, based on logistic regression, which has significant advantages over the area-based methodology. It does not require the choice of an arbitrary threshold of solvent ASA to define the contact-capable surface, it does not assume a linear dependency of contact frequency on ASA and it allows us to rationalize the propensities in terms of physicochemical properties such as charge, side-chain entropy and polarity.

Firstly, we derive a crystal contact-propensity scale for all 20 amino acids and we show that Gly and small hydrophobic residues top the list, with Glu and Lys having the lowest rank. Thus, crystal contacts are systematically depleted of residues with high side-chain entropy crystal contacts. This observation is consistent with the notion of anisotropic nonrandom protein–protein interactions in solution during crystallization, *i.e.* patch–patch interactions (Pellicane *et al.*, 2008). Furthermore, we also show that side-chain entropy rather than polarity is the key determining factor in these interactions. However, polarity appears to play a dominant role in the actual packing of larger contacts, so that apolar amino acids are systematically located towards the core of the contact.

Our results lend strong theoretical support to the concept of rational crystal engineering and specifically the surface-entropy reduction (SER) strategy (Derewenda, 2004; Derewenda & Vekilov, 2006). The approach was originally suggested based on the simple theoretical premise that loss of conformational degrees of freedom by large side chains incorporated into crystal contacts constitutes a potentially critical impediment to protein crystallization (Longenecker *et al.*, 2001; Mateja *et al.*, 2002). Subsequently, we designed and implemented a server that identifies suitable mutation sites based on amino-acid sequence information alone (Goldschmidt *et al.*, 2007). Using the SER strategy, a significant number of proteins have been successfully crystallized and their structures solved both in our group (Derewenda *et al.*, 2004; Devedjiev *et al.*, 2004) as well as other laboratories (Levinson *et al.*, 2008; Yip *et al.*, 2005). The conclusions of this paper may help to further refine the surface-engineering strategies.

After this paper was submitted, a study analyzing the physical properties that control protein crystallization and based on large-scale experimental data was published by the Northeast Structural Genomics Consortium (Price *et al.*, 2009). The authors analyzed a sequence database of 679 proteins, of which 157 were crystallized, and used logistic regression to identify protein-sequence features that impact on the binary outcome of the crystallization effort. The study concluded that surface entropy dominates all other effects, so that the fractional content of amino acids in the target sequence can be used as predictive parameters for crystallization. The approach used in that work is different from ours in that it attempts to derive the propensity of proteins to form well diffracting crystals from global sequence features by comparison of crystallizable and noncrystallizable proteins, whereas our analysis focuses exclusively on surface properties in proteins of known structure. Nevertheless, the fact that different computational approaches lead to virtually identical conclusions is most encouraging.

This work was supported by the PSI2 Program funded by NIGMS (NIH): U54 GM074946-01. We thank Drs Tom Terwilliger (Los Alamos National Laboratory) and David R. Cooper (University of Virginia) for helpful critiques of the early drafts of this manuscript and the anonymous reviewers for insightful comments that were invaluable in the final revision.

References

Albou, L. P., Schwarz, B., Poch, O., Wurtz, J. M. & Moras, D. (2008). *Proteins*, doi:10.1002/prot.22301.
Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. (2004). *J. Mol. Biol.* **336**, 943–955.
Berman, H., Henrick, K., Nakamura, H. & Markley, J. L. (2007). *Nucleic Acids Res.* **35**, D301–D303.
Carugo, O. & Argos, P. (1997). *Protein Sci.* **6**, 2261–2263.
Cheung, J. K., Shen, V. K., Errington, J. R. & Truskett, T. M. (2007). *Biophys. J.* **92**, 4316–4324.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Cooper, D. R., Boczek, T., Grelewska, K., Pinkowska, M., Sikorska, M., Zawadzki, M. & Derewenda, Z. (2007). *Acta Cryst.* D**63**, 636–645.
Dasgupta, S., Iyer, G. H., Bryant, S. H., Lawrence, C. E. & Bell, J. A. (1997). *Proteins*, **28**, 494–514.
Derewenda, U., Mateja, A., Devedjiev, Y., Routzahn, K. M., Evdokimov, A. G., Derewenda, Z. S. & Waugh, D. S. (2004). *Structure*, **12**, 301–306.
Derewenda, Z. S. (2004). *Structure*, **12**, 529–535.
Derewenda, Z. S. & Vekilov, P. G. (2006). *Acta Cryst.* D**62**, 116–124.
Devedjiev, Y., Surendranath, Y., Derewenda, U., Gabrys, A., Cooper, D. R., Zhang, R. G., Lezondra, L., Joachimiak, A. & Derewenda, Z. S. (2004). *J. Mol. Biol.* **343**, 395–406.
Doig, A. J. & Sternberg, M. J. (1995). *Protein Sci.* **4**, 2247–2251.
Goldschmidt, L., Cooper, D. R., Derewenda, Z. S. & Eisenberg, D. (2007). *Protein Sci.* **16**, 1569–1576.
Jain, A. K., Murty, M. N. & Flynn, P. J. (1999). *ACM Comput. Surv.* **31**, 264–323.

Janin, J. (1997). *Nature Struct. Biol.* **4**, 973–974.
Janin, J., Bahadur, R. P. & Chakrabarti, P. (2008). *Q. Rev. Biophys.* **41**, 133–180.
Janin, J. & Rodier, F. (1995). *Proteins*, **23**, 580–587.
Jelsch, C., Longhi, S. & Cambillau, C. (1998). *Proteins*, **31**, 320–333.
Levinson, N. M., Seeliger, M. A., Cole, P. A. & Kuriyan, J. (2008). *Cell*, **134**, 124–134.
Levy, E. D. (2007). *Structure*, **15**, 1364–1367.
Li, W. & Godzik, A. (2006). *Bioinformatics*, **22**, 1658–1659.
Liu, H., Kumar, S. K. & Sciortino, F. (2007). *J. Chem. Phys.* **127**, 084902.
Liu, H., Kumar, S. K., Sciortino, F. & Evans, G. T. (2009). *J. Chem. Phys.* **130**, 044902.
Longenecker, K. L., Garrard, S. M., Sheffield, P. J. & Derewenda, Z. S. (2001). *Acta Cryst.* D**57**, 679–688.
Mateja, A., Devedjiev, Y., Krowarsch, D., Longenecker, K., Dauter, Z., Otlewski, J. & Derewenda, Z. S. (2002). *Acta Cryst.* D**58**, 1983–1991.
Moreira, I. S., Fernandes, P. A. & Ramos, M. J. (2007). *Proteins*, **68**, 803–812.
Negi, S. S. & Braun, W. (2007). *J. Mol. Model.* **13**, 1157–1167.
Pellicane, G., Smith, G. & Sarkisov, L. (2008). *Phys. Rev. Lett.* **101**, 248102.
Price, W. N. II *et al.* (2009). *Nature Biotechnol.* **27**, 51–57.
Saha, R. P., Bahadur, R. P., Pal, A., Mandal, S. & Chakrabarti, P. (2006). *BMC Struct. Biol.* **6**, 11.
Shiryayev, A., Li, X. & Gunton, J. D. (2006). *J. Chem. Phys.* **125**, 24902.
Soyer, A., Chomilier, J., Mornon, J. P., Jullien, R. & Sadoc, J. F. (2000). *Phys. Rev. Lett.* **85**, 3532–3535.
Trinquier, G. & Sanejouand, Y. H. (1998). *Protein Eng.* **11**, 153–169.
Wentzel, N. & Gunton, J. D. (2008). *J. Phys. Chem. B*, **112**, 7803–7809.
Yip, C. K., Kimbrough, T. G., Felise, H. B., Vuckovic, M., Thomas, N. A., Pfuetzner, R. A., Frey, E. A., Finlay, B. B., Miller, S. I. & Strynadka, N. C. (2005). *Nature (London)*, **435**, 702–707.
Zhu, H., Domingues, F. S., Sommer, I. & Lengauer, T. (2006). *BMC Bioinformatics*, **7**, 27.