doi:10.1016/j.jmb.2007.05.022 J. Mol. Biol. (2007) **372**, 774–797

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Inference of Macromolecular Assemblies from Crystalline State

Evgeny Krissinel* and Kim Henrick

European Bioinformatics Institute, Genome Campus Hinxton, Cambridge CB10 1SD, UK

**Corresponding author*

We discuss basic physical–chemical principles underlying the formation of stable macromolecular complexes, which in many cases are likely to be the biological units performing a certain physiological function. We also consider available theoretical approaches to the calculation of macromolecular affinity and entropy of complexation. The latter is shown to play an important role and make a major effect on complex size and symmetry. We develop a new method, based on chemical thermodynamics, for automatic detection of macromolecular assemblies in the Protein Data Bank (PDB) entries that are the results of X-ray diffraction experiments. As found, biological units may be recovered at 80–90% success rate, which makes X-ray crystallography an important source of experimental data on macromolecular complexes and protein–protein interactions. The method is implemented as a public WWW service†.

Keywords: macromolecular assembly; asymmetric unit; biological unit; protein–protein interactions

Introduction

Macromolecular assemblies are complexes of more than one polypeptide and/or nucleotide chain that are stable in the native environment. The way in which the chains assemble represents the protein quaternary structure (PQS). Often (but not always), an assembly is the biological unit that performs a certain physiological function by facilitating respective biochemical processes. The functionality of many, if not most, proteins is dependent of the context of a macromolecular assembly. A simple example is given by the two-gene product, hemoglobin.[1] This protein complex, made of four polypeptide chains, is responsible for oxygen transport in the body, while no functional significance may be assigned to the isolated chains. Other important classes of macromolecular assemblies include holoenzymes, ion channels, DNA polymerase, microtubules, nucleosomes, virons, and many others.[2]

The physiological function of macromolecular complexes is known to be closely related to their 3D structure. While various techniques (e.g., light-scattering,[3] X-ray and neutron scattering,[4] mass spectrometry[5]) have been developed to study different properties of macromolecular assemblies, such as molecular weight, accessible surface area, chemical composition, and others, inference on the 3D structure is difficult in such experimental studies. Certain conclusions about the shape of assembly may be derived from mobility and mass measurements,[3] as well as from experiments on small-angle scattering.[6] Electron microscopy (EM) is applicable to studying large complexes, but it offers only low-resolution images. About 20% of structures in Protein Data Bank (PDB[7]) were obtained using NMR technique,[8] which is capable of getting atomic coordinates of macromolecular complexes in a solution. However, this method has limitations on the size of objects under study and is hardly applicable to medium and large assemblies. Besides, macromolecular complexes often exist in dynamic equilibrium, which further complicates interpretation of experimental results.

More than 80% of PDB entries were obtained by means of X-ray diffraction on macromolecular crystals.[9] It is reasonable to expect that stable

† http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html

Abbreviations used: PQS, protein quaternary structure; PDB, Protein Data Bank; ASU, asymmetric unit; ASP, atomic solvation parameter; MD, molecular dynamics; PISA, Protein Interfaces, Surfaces and Assemblies.

E-mail address of the corresponding author: keb@ebi.ac.uk macromolecular complexes do not change during crystallization and therefore they should be identifiable in crystal packing. By convention, a PDB entry contains only the atomic coordinates for the asymmetric unit (ASU) of a crystal. ASU is defined as the smallest unit that can be rotated and translated to generate one unit cell, using only the symmetry operators allowed by the crystallographic symmetry. Generally speaking, ASU may be chosen in many different ways, from which any one that contains the crystallographically unique covalently linked structure(s) may be acceptable for PDB deposition. However, macromolecular complexes, as a rule, are linked by weaker, non-covalent, interactions, and often possess crystallographic symmetry. As a result, a macromolecular complex may be made of a single or several ASUs, or several parts of neighboring ASUs, or several complexes may be contained in a single ASU. The lack of a direct relationship between ASU and macromolecular complex poses considerable difficulties for the identification of the latter in crystal packing in a universal manner.

Inference of macromolecular assemblies from crystalline state is often seen as a bioinformatical problem. In the framework of informatics-based approaches, macromolecular interfaces, found in crystals, are classified into "biologically relevant" and "insignificant" (crystal packing) ones according to a certain scoring system (cf. e.g., Ponstingl *et al.*[10]). The score may depend on the interface area, residue/atom composition and contacts, hydropathy index, charge distribution, topological complementarity, and other parameters. Disengagement of "insignificant" interfaces breaks the crystal apart, hypothetically leaving monomeric chains assembled by "significant" interfaces into biological units. This idea has found two different technical implementations. The first one was the Protein Quaternary Structure (PQS) server at the Macromolecular Structure Database group of the European Bioinformatics Institute (EBI-MSD),[11] which builds assemblies by progressive addition of suitable chain contacts. Another approach is represented by PITA (Protein InTerfaces and Assemblies) software,[12] which starts with the largest complex allowed by crystal symmetry and then iteratively splits it by bisectioning until a chosen threshold score is achieved. The interface scores in PITA were calibrated in the course of an exhaustive study on statistical discrimination between crystal contacts formed by homodimeric and monomeric proteins.[10]

There are, however, grounds to believe that interface properties alone are not indicative enough for unambiguous discrimination between relevant interfaces and artifacts of crystal packing. Indeed, if the binding energy of a particular interface is sufficient for dimerization of given macromolecules, it does not necessarily mean that an identical interface will bind a pair of considerably heavier objects. This was implicitly confirmed in a detailed study of interface properties reported in Jones and Thornton.[13] It has been concluded that no ultimate discriminating parameters for the identification of biologically relevant protein interfaces may be proposed even in the simplest case of dimeric complexes and that assessment of interface biological significance should take assembly type into account. Many other attempts to assess the significance of protein interfaces have been performed,[14–26] but no universal criteria were found. A few databases of protein–protein interactions and interfaces, derived from PDB, have been developed[25–28] in attempt to provide a systematic view on the factors that are responsible for macromolecular binding. One can expect that such databases and statistical analysis of different interface properties may be useful for the identification of transient interactions, which are extremely specific to the topology and chemical composition of binding sites. However, formation of stable complexes involves an interplay between affinity and entropy change and therefore it may be (and in fact it has been found to be, as shown below) less dependent on the interface characteristic features.

In this study, we discuss physical–chemical principles of macromolecular complexation and assess complex stability from the positions of chemical thermodynamics. We show that the binding energy of an interface is not a sole function of interface properties but also depends on the complex size and shape. We also show that entropy change due to complex formation is a major factor that, along with binding properties, determines complex size and symmetry. Next, we find that available theoretical approaches to the calculation of binding energy and entropy of dissociation have sufficient accuracy for the correct identification of macromolecular assemblies in crystals in 80–90% of instances.

Macromolecular Complexes in Solutions

Complexes differ from molecules in that, typically, their subunits do not make strong (covalent) chemical bonds. Rather, formation of complexes is only partly due to immediate (contact-dependent and electrostatic) interactions between the subunits. The dominant factors determining complex size and geometry are due to interaction with the solvent, therefore, existence and stability of complexes cannot be considered out of solvent context. Because of the relatively weak binding of their subunits, complexes often exist in dynamic equilibrium between different multimeric forms, in which the equilibrium is subject to subunit concentrations, temperature, and solvent properties. In many instances, a macromolecular complex cannot be identified unambiguously. In this section, we give an interpretation of the complex stability used in our approach and discuss different factors that affect formation and dissociation of macromolecular complexes, relevant to the problem of their identification.

Stability of macromolecular complexes

Consider complex $A = (A_1, A_2 \ldots A_n)$ made of subunits A_i. Any subunit may be a complex or a monomeric unit, such as a protein or DNA/RNA chain or a ligand. Upon bringing subunits together into a complex, part of their surface becomes inaccessible to solvent. We will call buried surface between subunits A_i and A_j as interface I_{ij}.

Dissociation of complex A involves a change in the standard Gibbs free energy:[29]

$$\Delta G^0_{\mathrm{diss}} = -\Delta G_{\mathrm{int}} - T\Delta S \qquad (1)$$

where ΔG_{int} is binding energy of subunits A_i, T is absolute temperature, and ΔS represents entropy change upon dissociation. $\Delta G^0_{\mathrm{diss}}$ depends on the set of subunits $\{A_1, A_2 \ldots A_n\}$, which we will call as "dissociation pattern". In general, dissociation may proceed along different patterns if they correspond to close values of $\Delta G^0_{\mathrm{diss}}$ and there are no dynamic factors making one pattern more preferable to others. However, for simplicity we will always assume that only one pattern with lowest $\Delta G^0_{\mathrm{diss}}$ is realized. $\Delta G^0_{\mathrm{diss}}$ may be expressed in terms of the equilibrium dissociation constant K_{d}:[29]

$$\Delta G^0_{\mathrm{diss}} = -RT\log K_{\mathrm{d}} = -RT\log \frac{\prod_{i=1}^{n} [A_i^0]}{[A^0]} \qquad (2)$$

where $[A^0]$ and $[A_i^0]$ are equilibrium concentrations of complex A and subunits A_i in units of the standard-state concentrations (1 M), at the standard-state temperature (300 K), and atmospheric pressure. At negative $\Delta G^0_{\mathrm{diss}}$, the dissociation proceeds spontaneously. Therefore, one can define complex as stable in the standard state if $K_{\mathrm{d}}<1$ or if $\Delta G^0_{\mathrm{diss}}>0$. Note that this definition does not always mean that the equilibrium complex concentration is higher than subunit concentrations under normal physiological conditions. As may be found from Eq. (2), assuming that neither of the subunits is found in an excess amount, $[A^0]\geq[A_i^0]$ is attained at $\Delta G^0_{\mathrm{diss}}\geq-(n-1)$ RT $\log[A_i^0]$. In case of dimers ($n=2$), at $[A_{i0}]=$ 1 mM, this yields $\Delta G^0_{\mathrm{diss}}\geq 4.1$ kcal/mol. This figure appears to be less than typical values of $\Delta G^0_{\mathrm{diss}}$, therefore, in many cases the above definition of complex stability implies prevalence of complex concentration.

Macromolecular binding in solutions

Many factors are known to contribute into macromolecule binding.[13–17,30–33] ΔG_{int} originates from the change of solvation energy ΔG_{solv} and immediate (contact-dependent ΔG_{cont} and electrostatic ΔG_{es}[37]) interactions between the subunits:

$$\begin{aligned}\Delta G_{\mathrm{int}}(A) = \Delta G_{\mathrm{solv}}(A) - \sum_{i=1}^{n} \Delta G_{\mathrm{solv}}(A_i) \\ + \sum_{j>i} \Delta G_{\mathrm{cont}}(A_i,A_j) \\ + \sum_{j>i} \Delta G_{\mathrm{es}}(A_i,A_j) \qquad (3)\end{aligned}$$

Solvation energy of macromolecule A is given by the work required for replacing a cavity of solvent C with the macromolecule:

$$\begin{aligned}\Delta G_{\mathrm{solv}}(A) = \Delta G_{\mathrm{cont}}(A,V) - \Delta G_{\mathrm{cont}}(C,V) \\ + \Delta G_{\mathrm{es}}(A,V) - \Delta G_{\mathrm{es}}(C,V) \qquad (4)\end{aligned}$$

where V stands for bulk solvent and all energies represent differences between vacuum and solvent environments. In this equation, $\Delta G_{\mathrm{cont}}(A,V)$ and $\Delta G_{\mathrm{cont}}(C, V)$ include any sort of contact interactions, such as hydrogen bonding (including the entropy change arising from full or partial immobilization of solvent molecules) between bulk solvent V and macromolecule A or cavity C, respectively. One can reasonably assume that contact interactions are proportional to the cavity surface area σ_{A}. Electrostatic interactions $\Delta G_{\mathrm{es}}(A,V)$ and $\Delta G_{\mathrm{es}}(C,V)$ depend, strictly speaking, on the position of all charged atoms in the cavity, in the macromolecule, and in the rest of the solvent. However, solvents in biological systems are water-based and therefore have a high value of dielectric constant $\epsilon_{\mathrm{s}}\approx 78$. Coulomb interaction in water fades on a distance scale of 7 Å (as given by the Onsager length $r_{\mathrm{C}}=e^2/(4\pi\epsilon_0\epsilon_{\mathrm{s}}k_{\mathrm{B}}T)$), which is relatively short comparing to the dimensions of biological macromolecules. Therefore, we may assume that $\Delta G_{\mathrm{es}}(C,V)$ depends mainly on the cavity surface area as well. Inside a macromolecule, dielectric constant is relatively low: $\epsilon_{\mathrm{M}}\approx 4$. This means that electrostatic interactions inside macromolecules have a range of $r_{\mathrm{C}}\approx 140$ Å and, as a consequence, $\Delta G_{\mathrm{es}}(A,V)$ cannot be represented by a contact term. As a result:

$$\Delta G_{\mathrm{solv}}(A) \approx \omega\sigma_A + \Delta G_{\mathrm{es}}(A,V) \qquad (5)$$

This conclusion is supported by experimental measurements of solvation energies of saturated hydrocarbon atoms in water ($\Delta G_{\mathrm{es}}(A,V)\approx 0$[34–36]), which suggest the value of $\omega\approx 7$ cal/(mol Å^2). Contact interaction between subunits $\Delta G_{\mathrm{cont}}(A_i, A_j)$ (cf. Eq. (3)) is mostly due to the formation of hydrogen bonds, salt bridges, and disulfide bonds across the interface I_{ij}. $\Delta G_{\mathrm{cont}}(A_i,A_j)$ may be approximated as:

$$\Delta G_{\mathrm{cont}}(A_i,A_j) = N^{ij}_{\mathrm{hb}}E^0_{\mathrm{hb}} + N^{ij}_{\mathrm{sb}}E_{\mathrm{sb}} + N^{ij}_{\mathrm{db}}E_{\mathrm{db}} \qquad (6)$$

where N^{ij}_{hb}, N^{ij}_{sb}, and N^{ij}_{db} stand for the number of hydrogen bonds, salt bridges, and disulfide bonds between subunits A_i and A_j, and E^0_{hb}, E_{sb}, and E_{db} are the average free energy gains per corresponding bond. Experimental and theoretical studies suggest that $E^0_{\mathrm{hb}}\approx 2$–10 kcal/mol.[39] Therefore, given that an average protein interface has 5–10 hydrogen bonds per 1000 Å^2,[40,41] hydrogen bonds may appear as a major contributor into $\Delta G_{\mathrm{int}}(A)$. However, in reality their contribution is considerably less significant because potential hydrogen bonding partners in the interfaces become satisfied by hydrogen bonds to water upon dissociation of the complex. This is accounted for in Eq. (3) by the sum of terms ΔG_{cont} (A_i, V) substituted from Eq. (4). The only effect that remains here is the decreased entropy of solvent due to the loss of mobility by bound molecules. Estimations show a final contribution of about $E_{\mathrm{hb}} \approx 0.6$–1.5 kcal/mol per bond only.[42,43] Limited experimental data on the stabilization effect of salt bridges suggest that free energy contribution of a salt bridge is close to that of a hydrogen bond $E_{\mathrm{sb}} \approx 0.9$–1.25 kcal/mol;[44,45] however, salt bridges are much rarer in interfaces ($\approx$1 per 1000 Å^2).[41] A disulfide bond may contribute up to 2–8 kcal/mol[46–48] at a yet lower occurrence. Using Eqs. (5) and (6), one can represent the free energy of binding (3) as:

$$\Delta G_{\mathrm{int}}(A) = \sum_{j>i} (-2\omega\Delta\sigma_{ij} + N_{\mathrm{hb}}^{ij} E_{\mathrm{hb}} + N_{\mathrm{sb}}^{ij} E_{\mathrm{sb}} + N_{\mathrm{db}}^{ij} E_{\mathrm{db}}) + \Delta G_{\mathrm{es}}^{*}(A) \tag{7}$$

where $\Delta\sigma_{ij}$ stands for the interface area between subunits A_i and A_j and the sum runs over all interfaces. The expression under the sum sign depends only on the properties of the corresponding interface, I_{ij}. $\Delta G_{\mathrm{es}}^{*}$ accounts for all interactions that cannot be attributed to isolated interfaces:

$$\Delta G_{\mathrm{es}}{}^{*}(A) = \Delta G_{\mathrm{es}}(A,V) - \sum_{i} \Delta G_{\mathrm{es}}(A_i,V) + \sum_{j>i} \Delta G_{\mathrm{es}}(A_i,A_j) \tag{8}$$

The main reason why electrostatic interactions in Eq. (8) cannot be attributed to isolated interfaces is their dependence on the environment through a dielectric constant $\boldsymbol{\epsilon}$. For example, $\Delta G_{\mathrm{es}}(A_i, A_j)$ depends on the size of complex and position of units A_i and A_j in it. As follows from Eqs. (7) and (8), representation of binding energy $\Delta G_{\mathrm{int}}(A)$ as a sum of interface binding energies $\Delta G_{\mathrm{int}}(I_{ij})$ results in improper description of electrostatic interactions. It is difficult to estimate the scale of the error involved without detail numerical calculations. A number of techniques for the calculation of electrostatic interactions have been developed, which are all applicable only if atom charges are known.[38,49–61] However, charge distribution inside a macromolecule represents a genuine problem because it depends on the molecule's environment: ionic strength of the solution, contact with other molecules, and other factors. Given the uncertainty in charge distribution, a rigorous account of electrostatic interactions is hardly possible. Therefore, for practical reasons, we do not make explicit electrostatic calculations in this study and assume that binding energy is given by an expression similar to the first term in Eq. (7):

$$\Delta G_{\mathrm{int}}(A) \approx \sum_{j>i} \left(\sum_{k} \omega_k \Delta\sigma_{ij}^{k} + N_{\mathrm{hb}}^{ij} E_{\mathrm{hb}} + N_{\mathrm{sb}}^{ij} E_{\mathrm{sb}} + N_{\mathrm{db}}^{ij} E_{\mathrm{db}} \right) \tag{9}$$

where index k enumerates different atom types, such as carbon, oxygen, etc., $\Delta\sigma_{ij}^{k}$ stands for the area of subunits A_i and A_j made of atoms of kth type buried in interface I_{ij}, and ω_k is the atomic solvation parameter (ASP) for kth atom type. The concept of ASPs has been used in many studies.[62–69] As may be concluded from above, ASPs do not have a real physical meaning and should be viewed as empirical parameters allowing for a better approximation of electrostatic term $\Delta G_{\mathrm{es}}^{*}$ in Eq. (7). Precision of solvation energy estimates with different ASP models was found to be within a few kilocalorie per mole,[63,67,69] which is close to that obtainable with electrostatic models.[54,70]

Entropy of complex dissociation

Calculation of absolute entropy S is a challenging task, which is not fully solved as of today.[71–75] Entropy of subunit A may be represented as:

$$S(A) = S_{\mathrm{trans}}(A) + S_{\mathrm{rot}}(A) + S_{\mathrm{vib}}(A) + S_{\mathrm{surf}}(A) \tag{10}$$

where S_{trans} and S_{rot} stand for the rigid-body translational and rotational entropy terms, respectively, S_{vib} is entropy of internal vibrational modes, and S_{surf} accounts for the entropy of surface atoms with fractional degrees of freedom.

Translational entropy term S_{trans} may be estimated using the Sackur–Tetrode equation, originally derived for classical ideal gas:[71–74]

$$S_{\mathrm{trans}}(A) = R\log\left[\left(\frac{2\pi m(A) k_{\mathrm{B}} T}{h^2} \right)^{3/2} \left(v e^{5/2} \right) \right] \tag{11}$$

where $m(A)$ stands for the molecular mass of subunit A, h is the Plank's constant, e is the Euler's number, and v denotes the free volume of solvent. The latter was introduced in Mammen *et al.*[74] in an attempt to account for the reduction of phase space in condensed media, where the volume of molecules itself cannot be neglected. As shown in Mammen *et al.*,[74] using an appropriate function for v allows one to calculate S as precise as 2% of the observed values in liquids, yet the parameter remains largely of empiric nature.

Rotational entropy term $S_{\mathrm{rot}}(A)$ can be calculated as logarithm of phase volume attainable by the rotational motion of A. The resulting expression reads:[71,73]

$$S_{\mathrm{rot}}(A) = R\log\left[\sqrt{\frac{\pi}{\gamma(A)}} \left(\frac{8\pi^2 k_{\mathrm{B}} T e}{h^2} \right)^{3/2} \sqrt{J_1(A) J_2(A) J_3(A)} \right] \tag{12}$$

where J_1, J_2, and J_3 stand for the principle moments of inertia of unit A. $\gamma(A)$ represents the symmetry number, defined as the number of unique rotations superposing A over itself. Comparison of Eqs. (11) and (12) indicates that $1/\gamma$ plays the role of free volume in rotational motion, which depends neither on the concentration $[A]$ nor the structure of solvent molecules. This is a probable reason for rotational entropies in liquids differ by only 2% of gas phase values[74] and Eq. (12) is found to be a good approximation in both gases and liquids.

Vibrational entropy S_{vib} may be estimated as a sum of vibrational entropies S^k_{vib} for all frequencies λ_k in the molecule's vibration spectra. S^k_{vib} can be calculated as in Einstein solid,[76] which finally gives:

$$S_{\mathrm{vib}} = R\sum_k \left[\frac{\beta_k}{e^{\beta_k}-1} - \log\left(1-e^{-\beta_k}\right)\right], \beta_k = \frac{hc\lambda_k}{k_{\mathrm{B}}T} \qquad (13)$$

where c stands for the speed of light. It is generally assumed that most vibrational modes of monomeric protein structures are not likely to be significantly affected by assembly,[74] therefore these modes are not expected to make a substantial contribution into entropy change ΔS. However, new modes emerge in complexes in place of degrees of freedom that get restricted by association. Depending on complex geometry and interface properties, these modes may be found in the low-frequency part of the spectra, where $\beta_k \ll 1$, and therefore they may contribute significantly into ΔS. This corresponds to the situation where large movements of associated subunits are allowed, which may be interpreted as only partial loss of their translational and rotational degrees of freedom upon assembly. This implies that weakly bound complexes may be stabilized by absorbing entropy in low-frequency vibrational modes. Estimates by Finkelstein and Janin[75] suggest that up to a half of $S_{\mathrm{trans}}+S_{\mathrm{rot}}$ may be transformed into S_{vib}; however, recent molecular dynamics (MD) studies showed a considerably smaller effect.[77] Calculation of vibration spectra $\{\lambda_k\}$ for large asymmetric structures, like proteins, is a computationally hard procedure, which involves many empirical parameters and functions. To our knowledge, no method can provide reliable estimates of λ_k in arbitrary case, and for this reason we neglect vibrational entropy in our analysis.

The last entropy term in Eq. (10), $S_{\mathrm{surf}}(A)$, is associated with the mobility of surface (side-chain) atoms of macromolecules. In first approximation, this term may be considered as proportional to the surface area of structure A:

$$S_{\mathrm{surf}}(A) = F\sigma_{\mathrm{A}} \qquad (14)$$

Substituting Eqs. (11), (12), and (14) into Eq. (10), and neglecting the vibrational entropy term S_{vib}, we obtain:

$$S(A) = C + \frac{3}{2}R\log(m(A)) + \frac{1}{2}R\log\left(\frac{J_1(A)J_2(A)J_3(A)}{\gamma^2(A)}\right) + F\sigma_{\mathrm{A}} \qquad (15)$$

where we introduce a constant parameter C, which we consider as an empirical one, even though analytical expression for it follows directly from Eqs. (11) and (12). There are several reasons to leave C as a free parameter in further analysis. Firstly, it depends on the empirically defined free volume v (cf. Eq. (11)). Secondly, C depends on the concentration $[A]$, which we would like to exclude from the consideration in an attempt to keep the software as simple for use as possible. As was pointed out in Stability of macromolecular complexes, our definition of complex stability refers to the standard state with $[A]=1$ M, which figure may be viewed as hidden in a particular value of C. If necessary, a concentration correction may be done using the following equation:[29]

$$\Delta G_{\mathrm{diss}} = \Delta G^0_{\mathrm{diss}} + RT\log\frac{\Pi_i[A_i]}{[A]} \qquad (16)$$

where second summand represents classical (zero own volume) dependence of entropy change $T\Delta S$ on the concentrations. Finally, we acknowledge that the described entropy models are simplified, and some entropy contributions such as vibrational and conformational entropies are not taken into account. Leaving C as an adjustable parameter may hopefully compensate, at least partially, for these factors.

Using Eq. (15), we obtain entropy change at complex dissociation:

$$\begin{aligned}\Delta S &= \sum_i S(A_i) - S(A) \\ &= (n-1)C + \Delta S_{\mathrm{trans}} + \Delta S^J_{\mathrm{rot}} + \Delta S^{\gamma}_{\mathrm{rot}} \\ &\quad + 2F\sum_{j>i}\Delta\sigma_{ij}{}^* \end{aligned} \qquad (17)$$

$$\Delta S_{\mathrm{trans}} = \frac{3}{2}R\log\left(\frac{\Pi_i\, m(A_i)}{m(A)}\right) \qquad (18)$$

$$\Delta S^J_{\mathrm{rot}} = \frac{1}{2}R\log\left(\frac{\Pi_{i;k}\, J_k(A_i)}{\Pi_k\, J_k(A)}\right) \qquad (19)$$

$$\Delta S^{\gamma}_{\mathrm{rot}} = R\log\left(\frac{\gamma(A)}{\Pi_i\, \gamma(A_i)}\right) \qquad (20)$$

Expression (17) cannot be represented as a sum over interfaces I_{ij}, which stands for the fact that entropy of dissociation is not a function of individual interfaces. Together with a similar conclusion, made in Macromolecular binding in solutions, this indicates that conventional bioinformatic approaches, based on interface scoring, cannot in principle provide a comprehensive description of macromolecular complexation.

It is well known that entropy drives thermodynamic systems to the most disordered (dissolved) state possible.[71] It is interesting to see that entropy also drives macromolecular complexes to less symmetric states. If a complex is formed by subunits with higher, then complex's, symmetry numbers $\gamma(A_i) \geq \gamma(A)$ (as in Figure 1(a)), $\Delta S^{\gamma}_{\mathrm{rot}}$ is negative and makes positive contribution into free energy of dissociation $\Delta G^0_{\mathrm{diss}}$ (cf. Eq. (1)). This corresponds to symmetry-assisted complexation. If the same complex is made of asymmetric subunits with $\gamma(A_i)=1$ (as in Figure 1(b)), $\Delta S^{\mathrm{rot}\gamma}$ becomes positive, which assists dissociation. Any rearrangement of subunits that decrease complex symmetry (see Figure 1(c)) results in increasing complex stability due to growth of $\Delta S^{\gamma}_{\mathrm{rot}}$ provided that dissociation pattern, binding energy, and com-

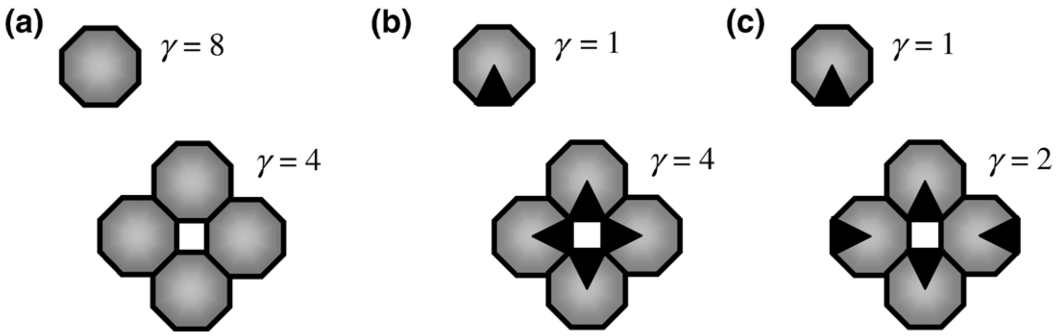

Figure 1. Symmetry effect on the formation of macromolecular complexes. For simplicity, 2D space is implied in this figure. (a) Formation of complex with symmetry number $\gamma(A)=4$ from four subunits with symmetry numbers $\gamma(A_i)=8$, each is assisted by negative entropy change $\Delta S^{\gamma}_{\text{rot}}$ (cf. Eq. (20)). (b) Formation of the same complex from subunits with symmetry numbers $\gamma(A_i)=1$, each is inhibited by positive $\Delta S^{\text{rot}\gamma}$. (c) Example of a less symmetric complex, which may be preferable to the one in (b), see discussion in the text. Note that, due to symmetry, dissociation patterns of all three complexes ought to be the same (four original subunits).

plex's moment of inertia do not change. In practice, however, ΔG_{int} usually reaches its minimum in the most symmetric packings and overweights the entropy drive to lower symmetries. Also, dissociation patterns usually depend on subunit packing. As a result, homomeric complexes are rarely asymmetric.

The symmetry effect on entropy of dissociation has analogy with the celebrated Gibbs paradox[78] in that the entropy change due to the difference of subunit orientations in Figure 1(b) and c does not depend on the nature and properties of subunits, on their interaction, and on what makes them asymmetric. The essence of the paradox is that $\Delta S^{\gamma}_{\text{rot}}$ does not change gradually with changing of any measure of asymmetry, such as r.m.s.d. or number of different atoms, but shows discontinuity at the point where subunits may be considered symmetric. It then appears that some complexes may be rated as stable or unstable depending on the chosen measure of asymmetry. The "paradox" is resolved in the same way as the classical one. As pointed out by Jaynes,[78] thermodynamic entropy is a property not of any one microstate (complex with a particular set of subunit orientations), but rather of a manifold of microstates, satisfying to the set of macroscopic quantities observed in a particular experiment. In the context of this study, it means that one should discriminate between complexes with different subunit orientations only if there is a physical process allowing to separate them by doing a measurable amount of work, and only if such a process is relevant to the specific conditions, in which the complexes are to be considered. In the next section, we list empirical criteria used by us for symmetry assessments. One can read more details about the Gibbs paradox and translate them onto subject in question from a beautiful discussion in Jaynes.[78]

Identification of Macromolecular Assemblies in Crystals

Theoretical analysis, presented in Macromolecular Complexes in Solutions, can answer the question whether a particular macromolecular assembly is stable, that is, may appear in solution in noticeable concentration on comparison with that of subunits. Our next goal is to identify stable complexes in macromolecular crystals and suggest scoring of their chances to be biological units.

Graph-theoretical search for solutions

We base our approach on the assumption that macromolecular assemblies retain their structures in crystalline state. This may be reasonably expected, noting that, according to Eq. (2), equilibrium fraction of assembly product in solution, $f=[A^0]/[A_i^0]=[A_i^0]^{n-1}/K_{\text{d}}$, only increases with an increase of subunit concentrations $[A_i^0]$ in the course of crystallization. Hypothetically, if K_{d} is not sufficiently low, assemblies may get crystallized along with isolated subunits, or an alternative assembly may emerge at higher concentrations. Therefore, we do not assume that a crystal is made of one particular type of assemblies. Instead, we are looking for the sets of different stable assemblies, which fill all the crystal space in a systematic manner. Each such set represents a candidate solution to the problem in question.

A crystal may be equivalenced with a periodic graph, where vertices and edges correspond to monomeric units and interfaces between them, respectively. We shall call an interface *engaged* if it connects two monomeric units in an assembly, and *disengaged* otherwise. Then each set of assemblies is unambiguously identified by the set of engaged interfaces. Due to crystal symmetry, engaged interfaces must satisfy the following conditions.

(1) If an interface of a particular type is engaged, all other interfaces of the same type (i.e., those formed by equivalent monomeric units in the same relative position) are also engaged.
(2) An interface cannot be engaged if doing so results in assembly that contains equivalent monomeric units in parallel orientations.

In order to understand condition 2, note that all parallel monomeric units in crystals may be obtained by repeated translations from one original unit. This means that any two parallel units, found at a given relative position in crystals, are connected by the same configuration of interfaces and other units. Therefore, if two parallel units are found in one assembly, then, due to condition 1, the assembly ought to include all their translation mates in similar relative positions, that is, to be of infinite size. As a consequence of condition 2, assembly size cannot exceed the size of unit cell.

One can find all assembly sets in crystals by enumerating all possible interface engagements satisfying to the above conditions. This may be efficiently addressed by a recursive backtracking scheme, commonly used in graph-matching algorithms (cf. e.g., Krissinel and Henrick[79]). Starting from the first type of interfaces in the list, the algorithm performs recursive calls, each of which engages other types of interfaces and disengages them upon withdrawal to lower levels of recursion. This scheme automatically satisfies condition 1 above. Condition 2 is satisfied by direct examination of the interface connections between parallel monomeric units before engagement of each interface. If, as a result of engagement, two parallel units get connected by interfaces, the corresponding branch of the recursion tree is rejected.

Since each N_I interface types may be in either of two states, engaged or disengaged, the maximal number of assembly sets to be explored is 2^{N_I}. Although this number is significantly decreased by subtracting states not satisfying condition 2, in many cases the problem remains computationally intractable. The number of combinations may be further decreased in two ways.

Firstly, engagement of some interfaces may automatically induce engagement of others. For example, if interfaces I_{12} and I_{23} are engaged in trimer (A_1, A_2, A_3), interface I_{13} can be found only in engaged state as well. Checking for induced engagements on each level of recursion eliminates whole branches of the recursion tree and thus efficiently decreases the search space.

Secondly, the algorithm may terminate branches that definitely do not lead to sets of stable assemblies. This technique has analogy with early-stage terminations in graph-matching algorithms.[79] In context of this study, the algorithm may terminate search down a particular branch of the recursion tree if ΔG^0_{diss} cannot become positive on any of the subsequent recursion levels. ΔG^0_{diss} reaches maximum at minimal ΔS and ΔG_{int} (cf. Eq. (1)); let us estimate them. Suppose that on recursion level r the algorithm has generated a set of unstable assemblies, so that $\Delta G^r_{\text{int}}+T\ \Delta S^r>0$ for all of them. Here, ΔS^r corresponds to the dissociation into a state represented by a lower recursion level $r_*<r$. Dissociation of assemblies from level $r+1$ may be viewed as that to level r and then spontaneously to level r_*. Note that entropy change between levels r and $r+1$, ΔS^{r+1}_*, is non-negative because it represents a transition from a more ordered to a less ordered state. Since entropy change in linked reactions is additive, $\Delta S_{r+1}=\Delta S_{*r+1}+\Delta S_r\geq\Delta S_r$. The binding energy term $\Delta G^{r+1}_{\text{int}}$ cannot be less than ΔG^r_{int} plus sum of binding energies of all hydrophobic interfaces that can still be engaged on subsequent recursion levels. Hence, the termination condition is:

$$\Delta G^r_{\text{int}}+T\Delta S^r+\sum_{jk}\min(\Delta G_{\text{int}}(I_{jk}),0)\geq 0 \qquad (21)$$

where $\{jk\}$ enumerates all interfaces that may be engaged on recursion levels $r+1$ and above. Despite a very general nature of this estimate, we found that it works very efficiently, especially if interfaces are engaged in order of increasing $\Delta G_{\text{int}}(I_{jk})$. Having little effect in cases with less than 10 interface types, condition (21) becomes vital for performance if N_I is greater than 20.

Dissociation pattern

In order to infer on the chemical stability of an assembly, one needs to calculate the Gibbs free energy of dissociation ΔG^0_{diss}, using Eqs. (1), (9), and (17). This is possible only if dissociation pattern $\{A_i\}$ is known. As was discussed in Stability of macromolecular complexes, we assume that assemblies dissociate into stable subunits along a pattern with minimal ΔG^0_{diss}. In addition, from symmetry considerations, the dissociation pattern cannot contain an engaged interface if an equivalent interface is disengaged anywhere else in the crystal. Dissociation patterns may be found by a backtracking scheme similar to the one described in Graph-theoretical search for solutions for assembly search. Represent the assembly as a graph, where vertices and edges correspond to monomeric units and interfaces between them, respectively. Starting with all interfaces in engaged state, the backtracking scheme enumerates all possible dissociation patterns by recursive disengagement of interface types. The same procedure is applied to all subunits A_i of every pattern in order to check them for chemical stability. Note that it is enough to detect any one pattern with negative ΔG^0_{diss} in order to conclude that assembly or a subunit is unstable, which allows to terminate backtracking search on early stages. We consider that identification of probable dissociation patterns is an important by-product of our approach, which by itself may be useful in applied studies.

Treatment of ligands

Many macromolecular structures in PDB contain small molecules (ligands) attached to their surface or buried in interfaces. The ligands may be both of natural origin or additives used in the crystallization procedure. Quite often, a particular oligomeric state is a direct result of ligand interactions. For example, inter-chain interfaces in PDB entries 1Y4L[80] and 2NYR[81] are not specific, and formation of dimeric complexes is mostly due to the presence of Suramin molecules, which penetrate deeply into interfacing structures and literally bridges them. Another type of example is given by PDB entries 1JL5 and 1G9U, where cylindrical tetrameric structures emerge only in the presence of calcium ions.[82] In general, it is the protein–ligand interactions that underlie all research in drug design. Therefore, consideration of ligands within macromolecular complexes is an important task to address.

All the analysis and algorithmic approach described in the previous sections may be directly applied to ligands, such that every ligand is considered as any other monomeric unit. However, for many PDB entries this results in a large number of interface types, which makes graph-theoretical approaches to assembly search and choice of dissociation patterns computationally intractable. Another inconvenience of straightforward approach to the inclusion of ligands, as was indeed found in the practical part of our study, is that most dissociation patterns become a mere detachment of ligands from the macromolecular complex, and macromolecular dissociation is then never seen.

In order to cope with computational difficulties and keep focus on macromolecular interactions, we "fix" ligands to macromolecules by permanent engagement of the corresponding interfaces, whenever necessary and possible. From empirical consideration, ligands of a particular type get fixed if they are less than 40 non-hydrogen atoms in size and if there are more than two of them in the ASU. If a ligand interfaces with more than one macromolecular unit, the one with the lowest binding energy to the ligand is chosen. A ligand cannot be fixed if it forms crystallographically equivalent interfaces to two or more monomeric units. In this approach, ligands effectively become a permanent part of macromolecular units they are fixed to, and may be viewed as surface modifiers that make an effect on binding properties of macromolecular units. We did not observe any change in oligomeric state (assembly size, geometry and composition) when comparing results obtained with and without ligand fixing. However, dissociation patterns (and therefore dissociation barriers $\Delta G^0_{\mathrm{diss}}$) were different in almost every case, with ligands detaching from a more stable macromolecular assembly when fixing was not used.

Identification of biological units

It should be acknowledged that thermodynamical stability is an important clue for concluding on the biological significance and function of a particular assembly, but it cannot be equated with the definition of a biological unit. The latter, generally speaking, depends on the context of a particular process and situation. It seems likely, however, that, in most instances, when the functional role of macromolecular assembly is not directly connected with dissociation, the biological unit should be given by the largest stable assembly that may exist under given physiological conditions. One can also reasonably expect that most biological units should represent highly stable complexes with low dissociation constant K_{d}. Following the discussion in the beginning of Graph-theoretical search for solutions, such complexes are expected to crystallize unchanged and the assembly search should yield a solution with a single structure.

Because no conclusions on biological function of macromolecular assemblies may be algorithmically derived in the framework of the present study, we keep all sets of stable assemblies, found by graph-theoretical search, as potential answers. Following the above considerations, we sort the sets and assemblies within the sets in the following order of priorities.

(1) Larger assemblies take preference over smaller ones.
(2) Single-assembly sets take preference over multi-assembly sets.
(3) Assemblies with higher free energy of dissociation $\Delta G^0_{\mathrm{diss}}$ take preference over those with lower $\Delta G^0_{\mathrm{diss}}$.

To our experience, such sorting is most successful in bringing assemblies that have better chances to be biological units, on top of the list. From now on, whenever we refer to a predicted biological unit in this study, the topmost assembly in the thus sorted list of potential answers shall be meant, unless otherwise specified.

Implementation and calibration of parameters

Model equations for standard Gibbs free energy $\Delta G^0_{\mathrm{diss}}$, derived in Macromolecular Complexes in Solutions (Eqs. (1), (9), and (17)), include undefined values of ASPs ω_k, hydrogen bond, salt bridge, and disulfide bond contributions E_{hb}, E_{sb}, and E_{db}, respectively, as well as entropy parameters C and F. As discussed in Macromolecular Complexes in Solutions, no rigorous theoretical expressions may be suggested in order to estimate these parameters, therefore we choose them such as to achieve the best agreement with available experimental data on macromolecular complexes.

For the calibration of model parameters, we used previously published data set of 218 protein structures with experimentally verified multimeric states.[12] Atomic ASPs were chosen according to the scheme suggested by Eisenberg and McLachlan[63] which classes atoms into five types: carbon (C), sulfur (S), neutral oxygen and nitrogen (O/N), charged oxygen (O^-), and charged nitrogen (N^+). There is little agreement between ASPs suggested in different studies even for the same classification scheme of atom types (e.g., compare data from Eisenberg and McLachlan[63] and Wesson and Eisenberg[66]), therefore we felt free to consider ω_k as adjustable parameters as well.

The fitting procedure performs iterative calculations of multimeric states for all structures in the benchmark data set. This results in a number of candidate solutions, which include assemblies with correct multimeric states as well as those of bigger and smaller, than the correct one, sizes. On each iteration, model parameters were fitted such as to satisfy the following system of inequalities for as many entries in the data set as possible:

$$\begin{cases} \Delta G^0_{\text{diss}} > 0 \text{ for correct assemblies} \\ \Delta G^0_{\text{diss}} \leq 0 \text{ for all other assemblies of} \\ \qquad \text{same or bigger size} \end{cases} \tag{22}$$

Obtained parameters are then used for assembly calculations in the next iteration, and the process stops when the number of data set entries with correctly predicted multimeric states ceases to increase.

The same procedure was then used for the calibration of nucleic acid ASPs, using 212 protein–DNA complexes reviewed by Luscombe *et al.*[83] The nucleic acid atoms were classed into eight types: phosphorus (P), base ring nitrogen (N_{aro}), amide nitrogen (N_{ami}), carbonyl oxygen (O_C), phosphate oxygen (O_P), other oxygens (O_{na}), aliphatic carbon (C_{ali}), and other carbon atoms (C_{na}). Only nucleic acid ASPs were varied on this step, with protein ASPs and other parameters taken as a result of the previous fitting procedure. After fitting the nucleic acid ASPs, protein ASPs need to be revised in order to correct multimeric states for those protein–DNA complexes where predictions failed on the protein–protein interaction side, for which both data sets are merged and the final adjustment of all parameters is done.

Calculations of rotational entropy, dissociation patterns, and assembly search require identification of equivalent monomeric units and interfaces. The equivalence is detected by structural alignment, for which we used secondary-structure matching algorithm.[84] Two monomeric units were considered as equivalent if structure alignment yields the following values of secondary-structure matching's quality score Q and sequence identity SI:

$$\begin{aligned} Q &= \frac{N^2_{\text{align}}}{(1+(\text{RMSD}/3)^2)N_1N_2)} \geq 0.65 \\ \text{SI} &= \frac{N_{\text{ident}}}{N_{\text{align}}} \geq 0.9 \end{aligned} \tag{23}$$

where N_1 and N_2 stand for the number of residues in two structures, RMSD is r.m.s.d. between aligned C^{α}'s in a best-structure superposition, and N_{align} is the number of aligned residue pairs, of which N_{ident} pairs are formed by identical residues. The above thresholds on Q and SI are of a purely empirical nature. They were chosen by a gradual decrease of initial $Q=0.9$ and SI $=0.95$ in the course of manual examination of many results where correct predictions could not be obtained without equivalencing more remote structures. It should be noted, however, that $Q=0.65$ corresponds to high structure similarity, when superposed backbones show only a moderate deviation from each other. In a similar way, the following criteria for interface equivalence have been obtained: interfaces I_{ij} and I_{kl} are considered as equivalent if they are formed by equivalent monomeric units $A_i \simeq A_k$, $A_j \simeq A_l$, found in similar relative positions, such that superposition of A_i and A_k places A_j and A_l (and *vice versa*) in no more than 2.5 Å off their locations of best superposition.

The described approach to the identification of macromolecular assemblies in crystals has been implemented in software Protein Interfaces, Surfaces and Assemblies (PISA) and made available for public use as an interactive webserver‡. The server comprises a searchable database of pre-calculated results for all PDB structures solved by means of X-ray crystallography, and allows for upload of PDB and mmCIF coordinate files for interactive processing. The calculations are distributed over a variable number of CPU nodes (1.2 GHz Pentium 4), depending on task complexity. The latter depends mainly on the type and number of monomeric units in ASU, space symmetry group, and number of interface types. Typically, the calculation results are returned in less than 30 s, most difficult cases may take up to 20 min. The server also provides a detailed description of interfaces, structures and assemblies, their visualization, and database search tools.

Results and Discussion

Table 1 shows empirical parameters of theoretical models for binding energy ΔG_{int} (Eq. (9)) and entropy of dissociation ΔS (Eq. (17)), obtained by the fitting procedure described in Implementation and calibration of parameters. As found, the system of inequalities (22) remains underfit for the data sets used, which means that maximal number of satisfied inequalities is achieved by many different sets of fitted parameters. This implies that the data sets used may be insufficient for the calibration purposes, and it is possible that the results may be further improved by using a larger selection of data. Given many possible solutions, we have chosen one that yields amino acid ASPs closest to those previously published.[63] It should be acknowledged that there is little agreement on any specific values of ASPs in the literature. This is demonstrated by data given in columns A and B in Table 1, obtained in Eisenberg and McLachlan,[63] and a follow-up study,[66] respectively. The disagreement may be due to inconsistency of reference data, used for the fit, or it may be a consequence of the non-physical nature of ASPs, as pointed out in Macromolecular binding in solutions, or it may indicate that five ASPs leave the system underfit. However, in the course of our numerical experiments, we found that the results are not very sensitive to the particular

‡ http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html

Table 1. Empirical parameters entering Eqs. (9) and (17), obtained by fitting the multimeric states found in the benchmark set of 218 protein complexes from Ponstingl *et al.*[12] (1st and 2nd columns in amino acid part of the upper table) and 212 protein–DNA complexes from Luscombe *et al.*[83] (nucleic acid part of the upper table)

Amino acid ASPs (cal/(mol Å^2))				Nucleic acid ASPs (cal/(mol Å^2))	
		A	B		
C	16	16±2	4±3	P	−23
S	41	21±10	−17±22	O_P	53
O/N	−11	−6±4	−113±14	O_C	57
N^+	−37	−50±9	−169±31	O_{na}	23
O^-	−17	−24±10	−166±31	N_{aro}	0
				N_{ami}	−19
				C_{ali}	14
				C_{na}	−33

Entropy parameters	
$T\cdot C$ (kcal/mol)	−6.82
$T\cdot F$ (kcal/(mol Å^2))	10^{-4}
E_{hb} (kcal/mol)	0.44
E_{sb} (kcal/mol)	0.15
E_{db} (kcal/mol)	4.00

Value of $T\cdot C$ assumes that mass is measured in a.u. and distance in Å. Column A shows ASPs obtained by Eisenberg and McLachlan[63] and column B shows those by Wesson and Eisenberg.[66]

values of fitted parameters, as long as they satisfy the maximum number of inequalities (22).

As seen from Table 1, energy contributions of hydrogen bonds and salt bridges appear somewhat lower than the estimates found in the literature (cf. Macromolecular binding in solutions), however, within a reasonable range. Given that significant interfaces usually have 10–20 and more hydrogen bonds, their effect on binding appears to be comparable with that of hydrophobic interactions. Contribution from the side-chain mobility into entropy of dissociation, $T\cdot F$, is quite small, about 0.1 kcal/mol per 10^3 Å^2 of interface area. Nucleic acid ASPs, obtained by our fit, make RNA/DNA chains slightly hydrophobic, which agrees with their low solubility in water. The average opening energy of a base-pair amounts to approximately 2 kcal/mol, which is somewhat lower than found in MD studies (e.g., Giudice *et al.*[85] suggests 4–7 kcal/mol), but agrees reasonably well with other data (1.6 kcal/mol found in Daune[86]). The figure of ~2 kcal/mol per pair implies that hydrophobic interactions only assist the closure of nucleotide base-pairs, while major contribution comes from hydrogen bonding.

Tables 2 and 3 present the assembly classification results obtained for protein–protein and protein–DNA complexes used in the fit, respectively. Each row in the tables corresponds to a particular oligomeric class present in the corresponding set, and columns give the classification counts obtained for that class. Table 2 shows reasonably uniform success rates for all oligomeric classes, with a lowest value of 84% obtained for tetramers. Tetramers have been found as a less predictable class also in Ponstingl *et al.*[12] Comparison with benchmark results in Ponstingl *et al.*[12] shows an improvement of 5–8% in the classification success rate. On comparison, the PQS server at EBI-MSD[11] achieves 78% correct answers on the same data set. However, one should take into account that, unlike PISA and PITA, PQS parameters were never optimized for the given set of structures. As found, PQS and PISA agree with each other on the classification of 77% of all entries in the data set.

Classification of protein–DNA complexes shows a somewhat higher success rate (cf. Table 3), with a total of 93% correct predictions as compared to 90% achieved for protein–protein assemblies. However, results for dimers, pentamers, and decamers are not indicative because of a too low number of structures in these classes.

Figure 2 shows the distribution of misclassified biological units in Tables 2 and 3 over free energy of dissociation ΔG^0_{diss}. As seen from the figure, most of misclassifications are found in the region of ±5 kcal/mol. Most of the errors are probably due to the differences in experimental conditions, such as concentration, pH, salinity, and temperature at which the biological units are detected. Since we do not account for specific experimental conditions in our analysis, implicitly replacing them with "average physiological conditions" in the fitting procedure, misclassifications within ±5–10 kcal/mol should not be rated as unexpectable. However, as Figure 2 shows, there are few cases where error in ΔG^0_{diss} is considerably higher than could be reasonably expected even for relatively approximate models

Table 2. Assembly classification obtained for the benchmark set of 218 PDB entries, representing protein complexes, from Ponstingl *et al.*[12]

	1mer	2mer	3mer	4mer	6mer	Other	Sum	Correct (%)
1mer	49	3	0	1	1	1	55	89
2mer	3	71+11	0	2+1	0	0	76+12	93
3mer	1	0	22	0	1	0	24	92
4mer	2	2+1	0	26+6	0	1	31+7	84
6mer	0	0	0	0+1	10+2	0	10+3	92
						Total	196+22	90

The rows give counts of multimeric states obtained for assemblies annotated as monomeric, dimeric, trimeric, tetrameric, and hexameric in the benchmark set. Counts represented as $N+M$ stand for N homomers and M heteromers obtained, otherwise only homomers are listed.

Table 3. Assembly classification obtained for the benchmark set of 212 PDB entries, representing protein–DNA complexes, reviewed by Luscombe *et al.*[83]

	2mer	3mer	4mer	5mer	6mer	10mer	Other	Sum	Correct (%)
2mer	1	0	0	0	0	0	0	1	100
3mer	6	96	0	0	1	0	2	105	91
4mer	0	2	83	0	0	0	0	85	98
5mer	0	0	2	3	0	0	0	5	60
6mer	1	0	0	0	13	0	1	15	87
10mer	0	0	0	0	0	1	0	1	100
							Total	212	93

The rows give counts of multimeric states obtained for assemblies annotated as dimeric, trimeric, tetrameric, pentameric, hexameric, and decameric in the set.

of subunit interactions. Consider some of these cases in more detail.

PDB entry 1QEX (bacteriophage T4 gene product 9) shows the largest classification error of 106 kcal/mol in ΔG^0_{diss}. The predicted homohexamer (Figure 3(a)) dissociates into almost equally stable homotrimers (Figure 3(b), $\Delta G^0_{\text{diss}} \approx 90$ kcal/mol), which were identified as biological units in Rossmann *et al.*[87] One of biological functions of the trimers is to provide attachment of long-tail fibers to the virus baseplate. The unit connects to the baseplate at variable angles with three extended tails formed by N-terminal domains of the polypeptide chains, which was verified by EM studies.[87] One could imagine that N-terminal tails of 1QEX have high propensity to attachment, and therefore it is not surprising that engagement of two trimers with their N-terminal tails results in highly stable hexameric complexes. However, with this sort of explanation it is unclear why N-terminal tails do not interact with each other in the trimer and how association of trimers is avoided in the course of virus assembly. It appears that electron density maps of 1QEX allow for an alternative fit, which yields a different configuration of N-terminal tails, represented by PDB entry 1S2E[106] (private communication from Dr. Sergei Strelkov, University of Leuven, Belgium). As found in PISA analysis, 1S2E forms homotrimers that do not merge into hexamers (Figure 3(c)). We therefore conclude that the hexameric complex in Figure 3(a) is an artifact of inappropriate interpretation of electron density maps. This example shows that PISA may be used for choosing a most probable alternative, when different structure solutions emerge in a crystallographic experiment.

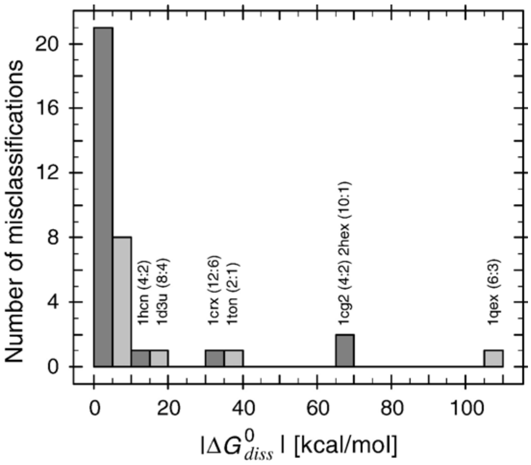

Figure 2. Distribution of misclassified biological units in Tables 2 and 3 over 5 kcal/mol intervals of free energy of dissociation ΔG^0_{diss}. In case of misclassification into a higher multimeric state, the classification error in ΔG^0_{diss} was defined as free energy of dissociation of that state. In case of misclassification into a lower multimeric state, the error was defined as negative free energy of dissociation of correct assembly. The labels give PDB codes of the largest-error misclassifications, where (N:M) denotes the misclassified (N) and correct (M) multimeric states.

Next to 1QEX in free energy classification error, PDB entry 1CG2 was predicted to be a homotetramer shown in Figure 4(a) instead of a dimer as in Figure 4(b). The tetramer is predicted to dissociate into the dimers at $\Delta G^0_{\text{diss}} \approx 66$ kcal/mol. This high dissociation barrier is mostly due to the presence of four Zn^{+2} in the catalytic domains, which make the binding between the homodimers. Removing Zn atoms from the file brings ΔG^0_{diss} down to 10 kcal/mol, which correlates reasonably well with the estimate of about 16 kcal/mol for Zn–protein binding, as reported in DiTusa *et al.*[89] No Zn binding is involved into the formation of homodimers, which are predicted to dissociate at ΔG^0_{diss} as low as ≈9 kcal/mol. Table 4 gives a summary of closest structural neighbors to PDB entry 1CG2 and their assembly classification by PISA software. *Q*-scores in the range of 0.2–0.4 indicate a moderate structure similarity, which is also reflected by low sequence identity. Sequence identity below 20% implies numerous residue substitutions on the protein surfaces, which is indeed confirmed by structure alignment. Extensive change in amino acid composition on the surface may have a pronounced effect on the binding properties of proteins and, as a result, cause a difference in crystal packing. Indeed, crystal packing of all structural neighbors in Table 4 does not contain the tetramer-forming interface equivalent to one in 1CG2; crystal packing of 1XMB does not include also the dimer interface. All structural neighbors to 1CG2 are therefore predicted to be dimeric, except 1XMB which does not exhibit complexation trends. Only two structures, 1FNO and 1Z2L, are predicted to form reliably stable homodimers, other structures appear to be on the

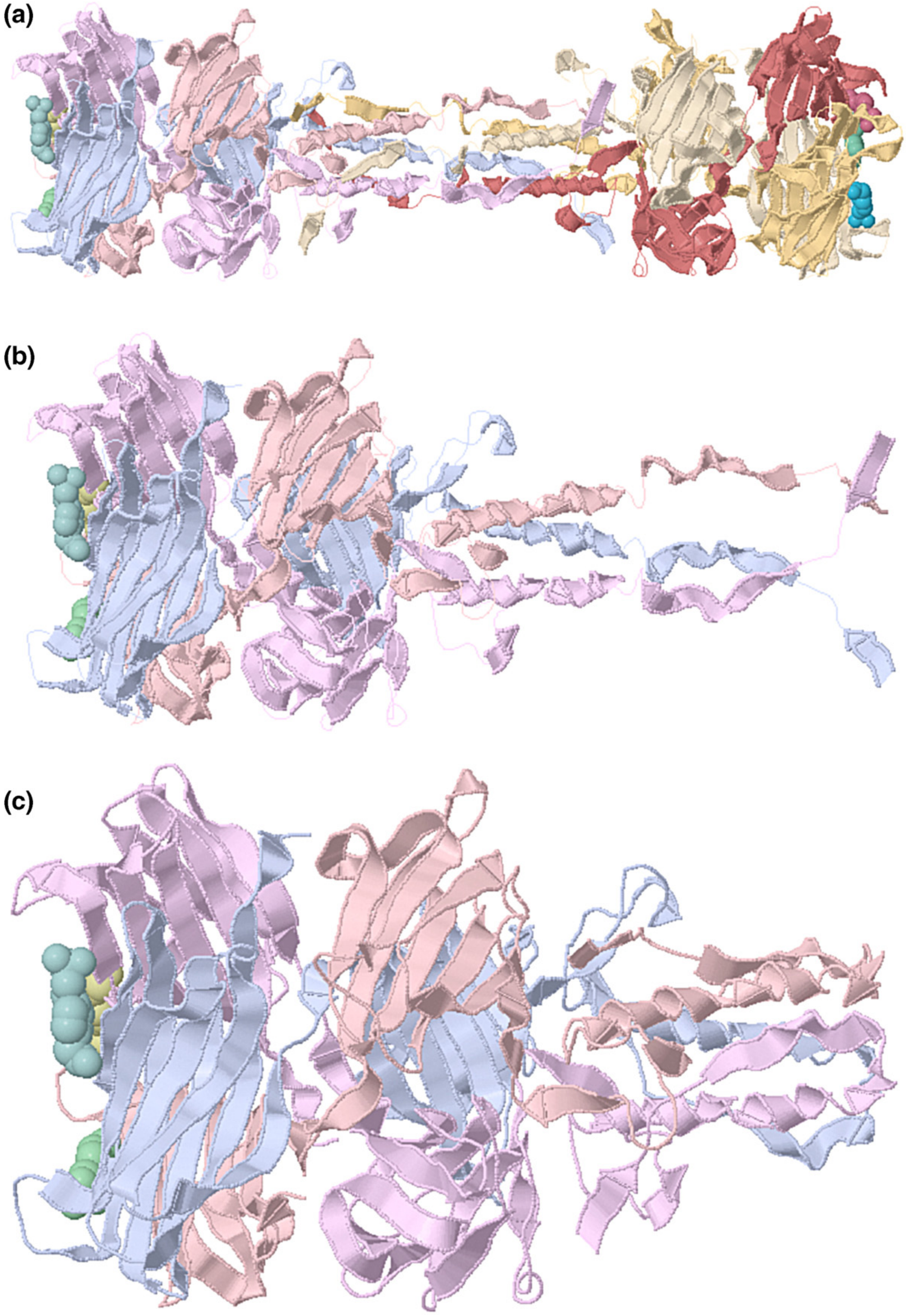

Figure 3. Homo-hexamer predicted for PDB entry 1QEX (a), the corresponding homo-trimeric biological unit (b) identified in Rossmann *et al.*,[87] and correct homo-trimer predicted for PDB entry 1S2E (c), representing an alternative fit of the same experimental data. See discussion in the text. The images were obtained using the Jmol software [http://www.jmol.org].

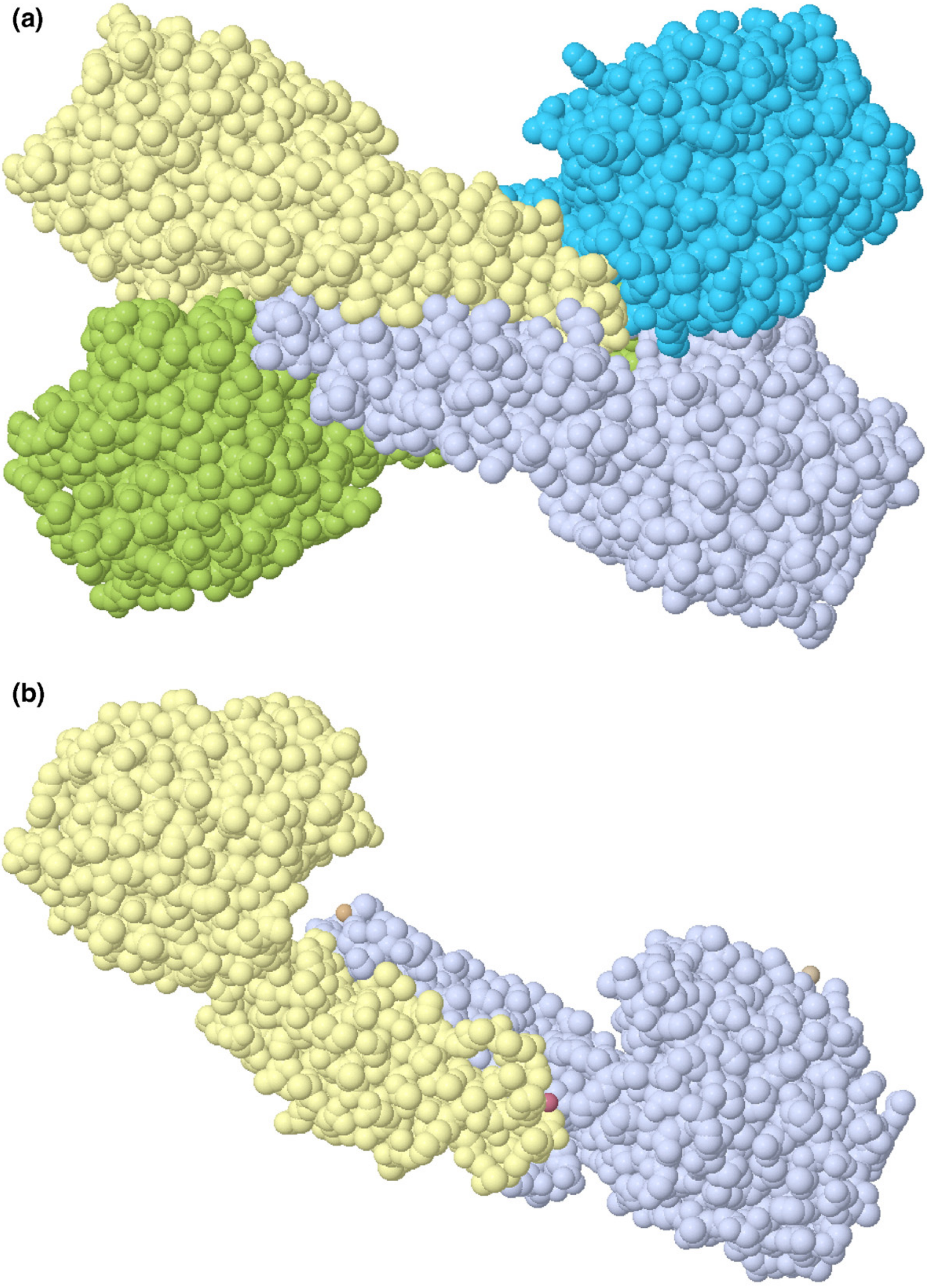

Figure 4. Homo-tetramer predicted for PDB entry 1CG2 (a), and the corresponding homo-dimeric biological unit identified in Rowsell *et al.*[88] See discussion in the text. The images were obtained using the Jmol software [http://www.jmol.org].

edge of stability. The presence of strong Zn-assisted interface between homodimers in 1CG2, not observed in crystal packings of its structural homologues, on one side, and experimental evidence from Rowsell *et al.*[88] that suggests a dimeric state, allows for the conclusion that misclassification of 1CG2 is due to a strong attachment of homodimers through their active sites in crystallization conditions.

Table 4. Closest structural neighbors to PDB entry 1CG2 and their assembly classifications, obtained by PISA

PDB code	Space group	ASU size	Assembly size	*Q*-score	Seq. Id	ΔG^0_{diss} (kcal/mol)
1CG2	P 1 21 1	4	4	1.00	1.00	66.2
1YSJ	C 1 2 1	2	2	0.38	0.16	3.9
1VGY	P 1 21 1	2	2	0.37	0.19	9.2
1XMB	P 32 2 1	1	1	0.35	0.18	—
1FNO	C 1 2 1	1	2	0.35	0.21	17.4
1VIX	P 21 21 21	2	2	0.34	0.20	8.3
1Z2L	P 21 21 2	2	2	0.20	0.14	34.4

Q-score is a measure of structural similarity (cf. Eq. (23)[84]), and sequence identity (Seq. Id) was calculated from structure alignment. The neighbors have been identified by structure search facility in PISA software based on secondary-structure matching algorithm.[84] See discussion in the text.

Emergence of interaction artifacts due to crystal packing appears to be the most common reason for the misidentification of biological units in crystalline state. All largest-error misclassifications in Figure 2, except for PDB entries 1QEX, 1D3U, 1CRX, and 1TON are due to this reason. 2HEX is predicted to be a strong homodecamer, dissociating into five homodimers at $\Delta G^0_{\text{diss}} \approx 65$ kcal/mol as shown in Figure 5(a) and (b). This decameric structure was noted in Lubkowski and Wlodawer[90]; however, the authors stated that "no biological relevance for BPTI decamers is known or postulated; therefore they must be considered the result of interactions in the crystal". In case of 1HCN, the predicted heterotetramer (shown in Figure 6) dissociates, at $\Delta G^0_{\text{diss}} \approx 15$ kcal/mol, into identical heterodimers, which were found to be biological units in Wu *et al.*[91] The dimers appear to be somewhat more stable than the tetramer ($\Delta G^0_{\text{diss}} \approx 20$ kcal/mol). Formation of the tetramer may possibly be assisted by the binding sites formed by residues 38–57 (shown in red in Figure 6), which were identified by antibody blocking in Lustbader *et al.*[92]

Results for PDB entry 1D3U (cf. Figure 7) exemplify a situation that is specific to protein–DNA complexes. The predicted octamer dissociates into two identical tetramers, identified as biological units in Littlefield *et al.*,[93] at $\Delta G^0_{\text{diss}} \approx 20$ kcal/mol. The tetramers, found as a second-choice solution in PISA output, make the left- and right-hand parts of the complex projection in Figure 7. In this case, PISA prediction fails because, strictly speaking, neither the crystallized octamer nor tetramer represents naturally found structures. For crystallization purposes, the virtually infinite DNA strands are replaced by chemically synthesized 24-base fragments, bound to protein parts of the complex. This replacement allows the tetramers to engage into a contact, which, contrary to basic PISA assumptions, cannot occur in natural conditions. As estimated by PISA, the interface between two contacting helices of the left- and right-side tetramers in Figure 7 shows a relatively high hydrophobic effect of $\Delta G_{\text{sol}} \approx -7.4$ kcal/mol and forms 18 hydrogen bonds and 8 salt bridges, which add a further ≈ -9 kcal/mol to the interface binding energy ΔG_{int}. In addition, as may also be seen in Figure 7, the DNA strands of the tetramers make a cross-pairing, which involves three bases from each side. This indicates a mutual affinity of the tetramers, which, in PISA estimates, would be strong enough for merging them into octamers if the inter-tetramer contacts were not a mere artifact of crystal packing.

From the first glance, the above considerations appear to be equally applicable to PDB entry 1CRX, which is predicted to be a hetero-dodecamer shown in Figure 8. The dodecamer is made of four very similar, but not all identical, hetero-trimers. Each of the trimers, considered as biological units,[94] includes a DNA fragment bound to bacteriophage recombinase Cre. The complex is predicted to dissociate into two hetero-hexamers, forming the upper and lower halves of the structure projection in Figure 8, at $\Delta G^0_{\text{diss}} \approx 28$ kcal/mol, although, from symmetry considerations, one would expect dissociation into trimeric state. A detailed examination reveals that dissociation into hexamers is due to asymmetry caused by the presence of modified *O*-phosphotyrosine residues in two out of four protein chains, shown in space-fill mode in Figure 8. The "vertical" protein–protein interfaces in Figure 8 are partly mediated by *O*-phosphotyrosine residues, which adds about −9 kcal/mol to their binding energies as compared with "horizontal" interfaces that disengage upon dissociation. This finding may imply that the dodecameric structure is not an artifact of crystal design as in the case of 1D3U in Figure 7. Indeed, the structure represents a synapsed complex performing site-specific DNA recombination.[94] The three-stage reaction opens two DNA strands, initially running across the upper and lower hexamers in Figure 8, and recombines them into strands running vertically through the left and right halves of the dodecameric structure in the figure. It is therefore obvious that the dodecameric complex represents a real intermediate structure, which performs a certain physiological function and for this reason could be classed as a biological unit. This case demonstrates a situation where definition of biological unit is rather subjective and cannot be automated.

The last out of the strongest misclassifications in Figure 2, PDB entry 1TON, is predicted to be a dimeric complex with a dissociation barrier of $\Delta G^0_{\text{diss}} \approx 37$ kcal/mol. Visual inspection reveals that dimerization is assisted by two Zn ions mediating the interface. Removal of Zn ions from the structure brings the dissociation barrier down to 3.2 kcal/mol, again in a good agreement with the abovementioned estimate of about 16 kcal/mol for Zn–protein binding.[89] The value of $\Delta G^0_{\text{diss}} \approx 3.2$ kcal/mol does not allow to reliably identify the structure as a monomer or dimer in view of finite precision of PISA models. In this particular case, Zn ions have been added to the buffer in order to aid crystallization,[95] therefore the predicted strong dimer

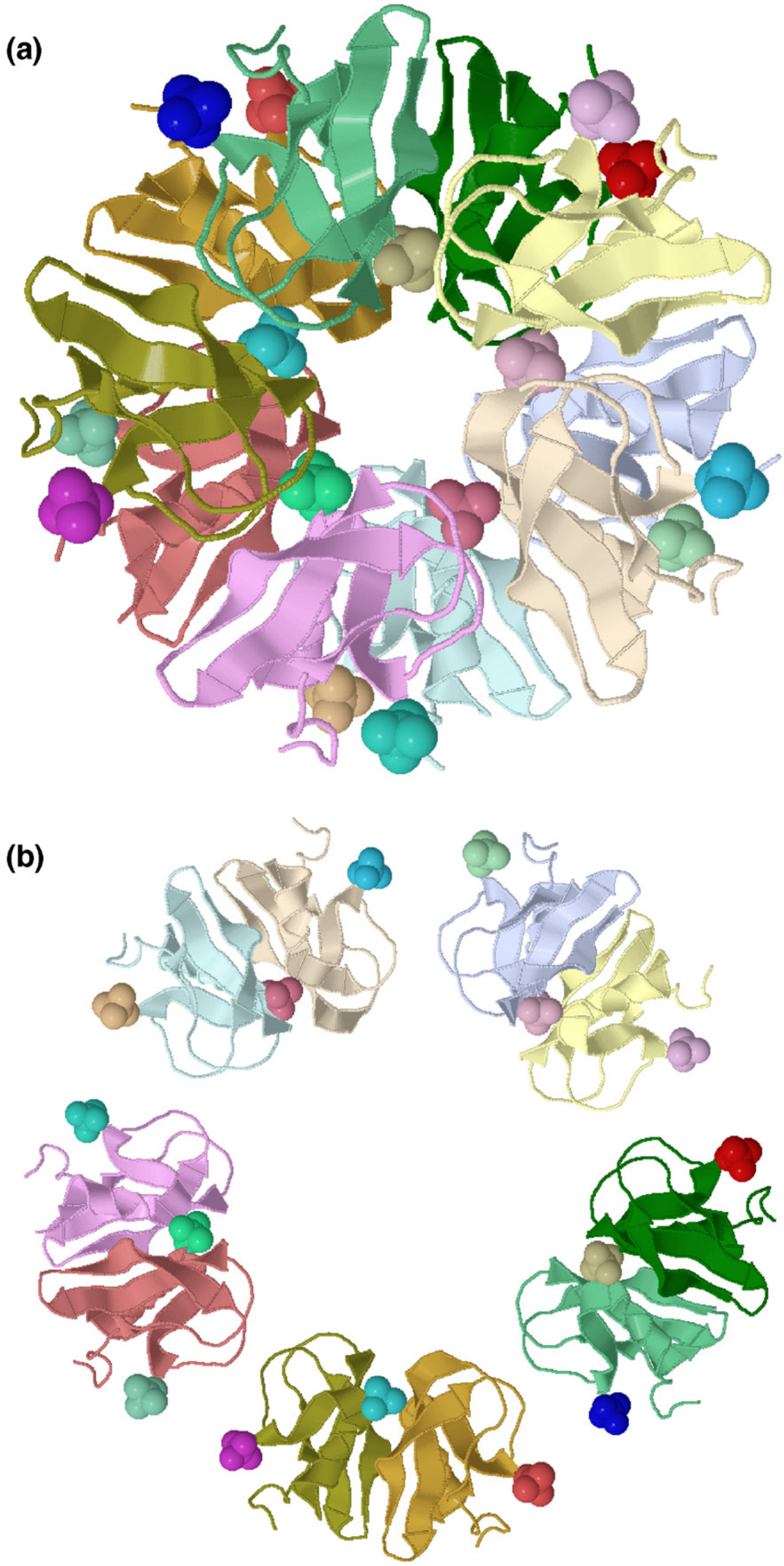

Figure 5. Homo-decamer predicted for PDB entry 2HEX (a), and its dissociation into 5 homo-dimers suggested by PISA analysis (b). The space-fill mode shows SO_4 molecules. See discussion in the text. The images were obtained using the Jmol software [http://www.jmol.org].

should be regarded as a clear artifact due to crystallization conditions. The presence of binding agents in crystals is a very common factor that affects automatic identification of biological units. Another bright example is given by PDB entries 1JL5 and 1GD9 (not included into the calibration data set), for which PISA predicts oligomerization into pipewise structures in the presence of Ca and Hg ions, and

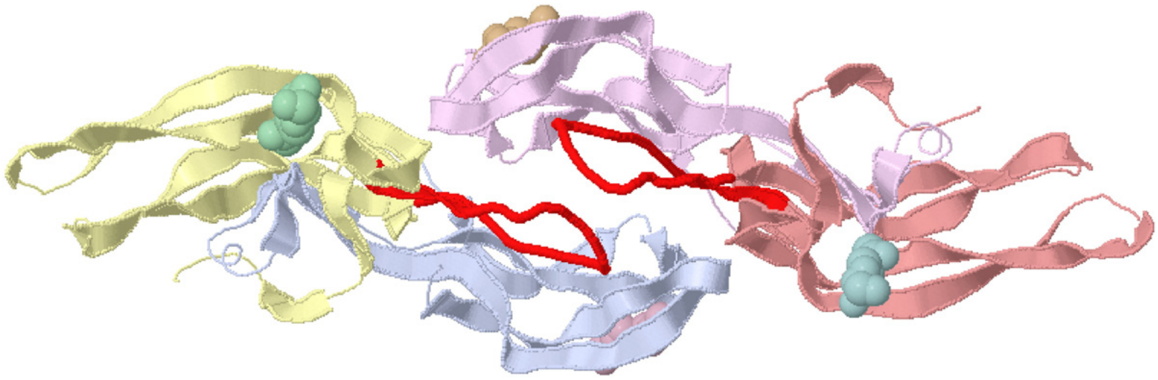

Figure 6. Hetero-tetramer predicted for PDB entry 1HCN, made of two identical hetero-dimers found to be biological units in Wu *et al.*[91] The space-fill mode shows NAG molecules, red color denotes residues identified by antibody blocking (cf. Lustbader *et al.*[92]). See discussion in the text. The image was obtained using the Jmol software [http://www.jmol.org].

monomeric states in their absence, in full agreement with experimental findings.[96]

The above analysis shows that all the strongest misclassifications within the data sets, used for calibration of model parameters, are caused by various factors that place the biological unit definition outside a simple concept of thermodynamically stable complex. In such cases, additional data should be used for correct classification, which, however, is difficult for algorithmic implementation. In other instances, when the definition of the biological unit may be reduced to the concept of stable complexes, successful classification is subject to the precision of PISA models for binding energy and entropy of dissociation. On the basis of misclassification results, shown in Figure 2, one could use the figure of ±5 kcal/mol as a classification margin in PISA. This figure can be hardly proven or disproven as the respective experimental data are not readily available. A limited user feedback, personal communications, and one recent analysis[97] suggest that the figure may be close to reality. Comparison of entropy change at dissociation of Fasciculin-2–Acetylcholinesterase complex,

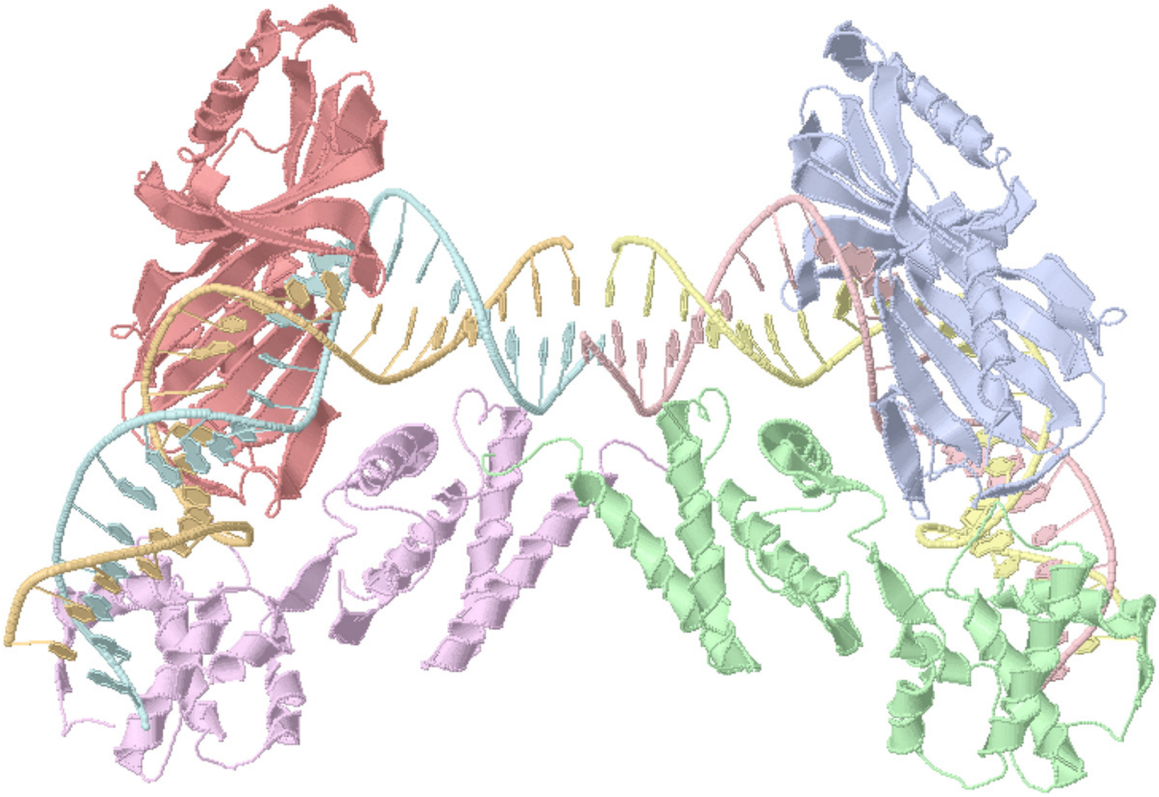

Figure 7. Hetero-octamer predicted for PDB entry 1D3U. The structure dissociates into two identical hetero-tetramers (the biological units, cf. Littlefield *et al.*[93]), making right- and left-hand parts of the shown complex. See discussion in the text. The image was obtained using the Jmol software [http://www.jmol.org].

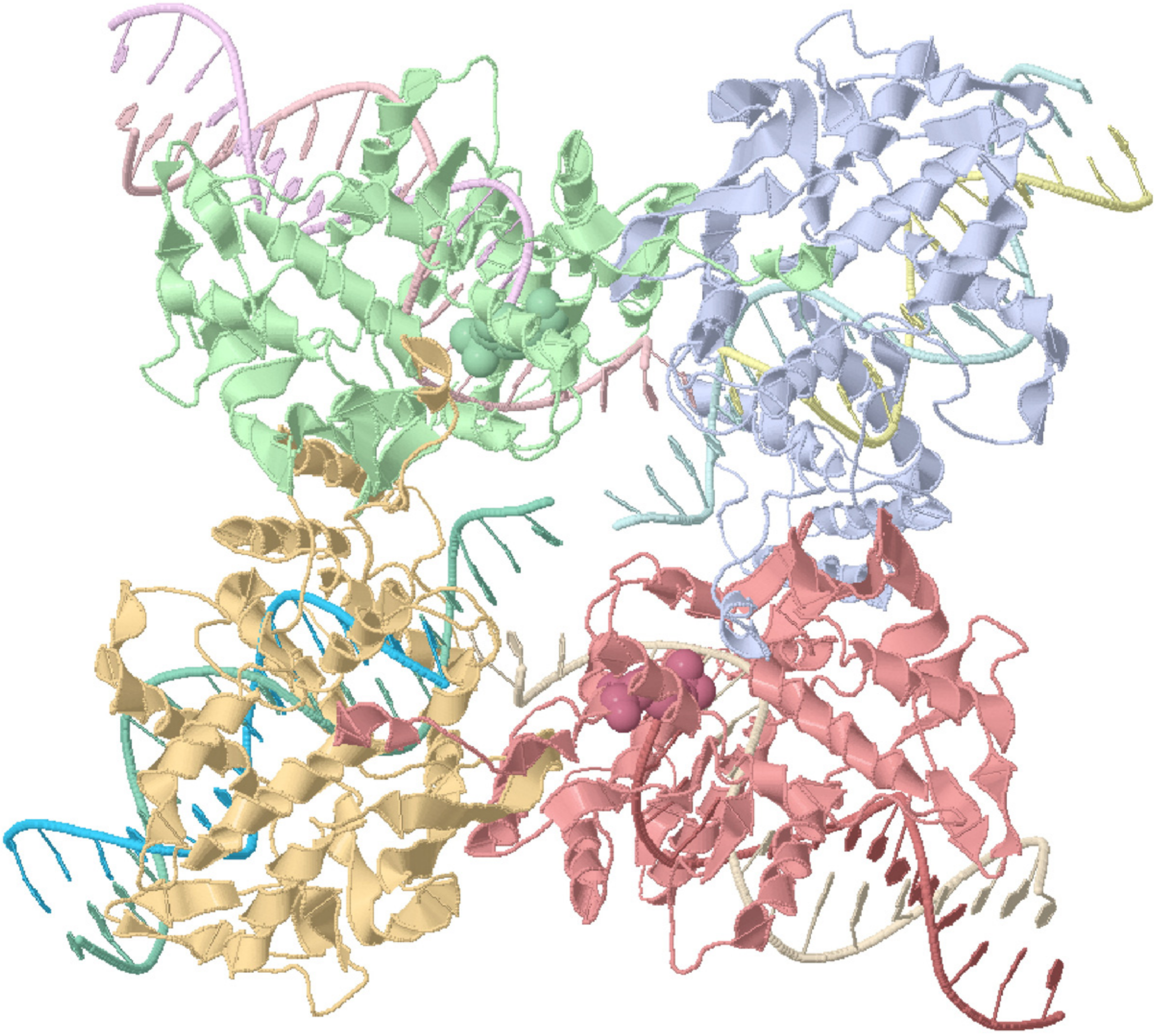

Figure 8. Hetero-dodecamer predicted for PDB entry 1CRX. The complex is made of four hetero-trimers, each including a DNA fragment bound to bacteriophage recombinase Cre, considered as primary units in Guo *et al.*[94] The space-fill mode shows *O*-phosphotyrosine residues found in two out of four protein chains. See discussion in the text. The image was obtained using the Jmol software [http://www.jmol.org].

estimated by PISA ($\Delta S \approx 37$ cal/mol K), and the one obtained in the course of molecular dynamic simulations ($\Delta S \approx 32$ cal/mol K),[77] results in only ≈ 1.5 kcal/mol difference in $\Delta G^0_{\rm diss}$ at room temperature. The only essential difference between PISA estimates and the referenced MD studies is the account of vibrational modes in the latter. As part of entropy is absorbed by the vibrational modes corresponding to the relative motion of complex's subunits, the entropy change in MD studies is expectably lower than in our models. However, as the comparison shows, the resulting difference seems to be less than the overall precision in case of tightly bound complexes. It should be acknowledged here that PISA may miss the stabilization effect of vibrational entropy in flexible complexes.

As follows from above, there is a natural limit on the accuracy of automatic identification of biological units in crystals, which is due to the difference in experimental and physiological conditions and involvement of additional data on complex function. This limit would be there even if coordinate data, physical–chemical models, and calculation techniques were perfect. Figure 9 demonstrates the scale of uncertainty in the biological unit assignments by comparison of the results obtained from MSD database[98] and PISA servers at EBI-MSD. Biological unit assignments in MSD are based on PQS[11] predictions followed by manual inspection. As seen from the figure, 29,837 entries out of $\approx$31,900 X-ray structures currently in the PDB have been compared. The rest was ignored as non-comparable, where PISA solution contained more than one assembly type or where coordinate data did not allow for PISA analysis (e.g., where only backbone atoms are present). As may be derived from Figure 9, 72% of structures appear to be monomeric or dimeric, where MSD and PISA agree in 81±1% and 74±2% of instances, respectively. In multimeric classes up to hexamers, MSD and PISA agree on 76.4% of classifications, which coincides with the figure obtained for the calibration data set above. A simple

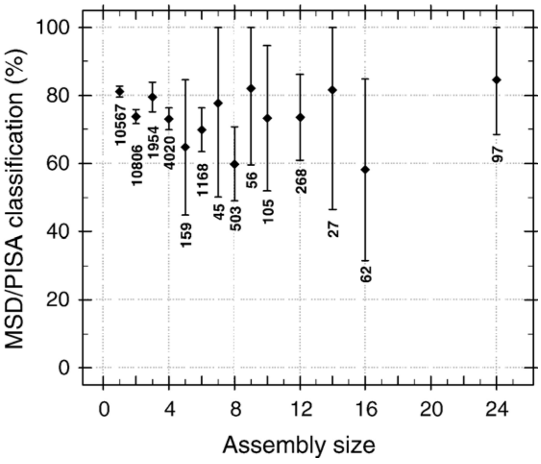

Figure 9. Comparison of multimeric state assignments obtained from MSD database[98] and PISA server at EBI-MSD. Diamonds denote percentage of coinciding assignments in the respective multimeric class, error bars indicate the 3σ confidence region. Only complete structures obtained by means of X-ray diffraction, giving single-assembly solutions in PISA have been used for the comparison and only multimeric classes up to 24mers with more than 20 structures are shown. Numbers below the diamonds give the total number of structures used for comparison in the respective multimeric classes [http://www.ebi.ac.uk/msd/].

average over first six multimeric classes with no regard to the actual number of classifications yields a very close figure of 73.7%. This may be taken as an implicit indication that the calibrated parameters work uniformly well on the whole PDB despite being fitted on just 1.35% of all entries.

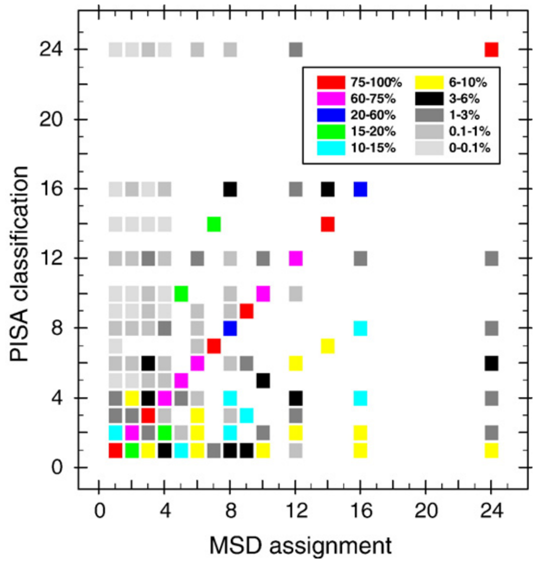

Figure 10. PISA classification of biological unit assignments in the EBI-MSD database.[98] Different colors denote percentage (as shown in the legend) of structures from a particular multimeric class in MSD, classified by PISA into the same or other classes. For example, 10–15% of structures, assigned monomeric state in MSD, were classified by PISA into dimers (cyan rectangle at (1, 2)). The distribution was calculated on the same data as that used for Figure 9.

Figure 10 shows more details on the relation between biological unit assignments in MSD and PISA. As seen from the figure, PISA tends to classify structures into lower oligomeric states, as compared with PQS/MSD. A possible explanation to this fact is that PQS predictions are largely based on the per-chain average scores for buried surface area, solvation energy gain, hydrogen bonds, and others. Therefore PQS may allow for a relatively weak interfaces between subunits, if their scores are compensated by scores of strong interfaces. This is partly confirmed by the observation that PISA assemblies tend to be subunits of MSD assemblies, where they are different. Indeed, as seen from Figure 10, most of the "misclassified" multimeric states in PISA represent, together with the MSD-assigned states, geometrical series. For example, hexadecamers are most often "misclassified" into octamers and tetramers, then dimers and monomers; nonamers

Table 5. Comparison of multimeric states predicted by PISA with those observed experimentally for a set of structures related to bacteriophage T4 studies

PDB	Multimeric state		
code	PISA	Observed	Ref.
1G31	14	14	99
1VQ2	6	2	100
1TEO	6	2	100
1K28	6	6	101
1H6W	6	6	102
1M5R	3	3	103
1IXY	3	3	103
1EL6	3	3	104
1OCY	3	3	105
1S2E	3	3	106
1CZD	3	3	107
1V1H	3	3	108
1V1I	3	3	108
2FKK	3	3	109
1N7Z	2	2	110
1N8B	2	2	110
1N80	2	2	110
1E7L	2	2	111
1E7D	2	2	111
1EN7	2	2	112
1BJA	2	2	113
1C1K	1	1	114
1C3J	1	1	115
1QKJ	1	1	115
1T8G	1	1	116
1T8F	1	1	116
1SSY	1	1	116
1SSW	1	1	116
1RIF	1	1	117
1NZF	1	1	118
1NZD	1	1	118
1NVK	1	1	118
1J39	1	1	118
1JEJ	1	1	119
1JG6	1	1	119
1JG7	1	1	119
1JIU	1	1	119
1JIX	1	1	119
1JIV	1	1	119
1L60	1	1	120
1KAF	1	1	121

Table 6. Comparison of multimeric states assigned by PISA, MSD, and PDB to structures representing the strongest PISA misclassifications from Figure 2

PDB code	Multimeric state				
	Literature	PISA	MSD	PDB(a)	PDB(350)
1QEX	3	6	6	3	4
1CG2	2	4	4	2	Not given
2HEX	1	10	10	Not given	5
1HCN	2	4	4	Not given	Not given
1D3U	4	8	8	Not given	4
1CRX	3	12	12	2	6
1TON	1	2	2	Not given	Not given

Literature sources (cf. above) give experimentally verified states. PDB(a) corresponds to biological unit annotation in PDB files, PDB(350) refers to multimeric states inferred from PDB remark 350.

may appear in PISA as trimers, heptamers as tetradecamers, and so on.

As was mentioned before, a very limited number of PDB depositions come with experimentally verified multimeric states. One of the best studied classes of structures in this respect is represented by the series of works on bacteriophage T4. Table 5 compares PISA predictions for these structures with experimentally identified multimeric states. The PDB entries, listed in the table, were obtained by keyword search on "phage T4" in PISA database, from where we excluded entries without experimental evidence for multimeric states and two entries (1QEX and 1AA0) used in the calibration data set. We also excluded PDB entry 1MVA (T4 capsid), which is not good for PISA analysis as it does not allow for reconstruction of the ASU because of the absence of Non-Crystallographic Symmetry (NCS) records. As may be seen from the table, PISA analysis failed only in two instances (1VQ2 and 1TEO), where hexameric states were predicted instead of observed dimers. It is interesting to note that these entries represent dimeric mutants of a wild protein that has a natural hexameric state.[100] The mutants crystallize as hexamers, which means that PISA predictions in these cases are not too far from the reality. Overall accuracy of PISA predictions, shown in Table 5, is just under 95%.

We would like to point out here that despite enormous curation effort, neither MSD nor PDB should be rated as a source of validated data on quaternary structures. Table 6 gives a summary of multimeric states for structures representing the strongest PISA misclassifications from Figure 2, discussed above in detail. As seen from the table, MSD assignments coincide with those of PISA in these particular cases and are all wrong as well, while the results inferred from the corresponding PDB files look confusing and far from complete. It is therefore not possible to derive any definite conclusions about quality of either source of data and trustworthiness of PISA predictions in general. We, however, find it encouraging that MSD and PISA assignments, which are both algorithm-based, agree on considerable (76%) fraction of available data. As has been mentioned before, MSD predictions are based on PQS,[11] which in its essence is a bioinformatic, interface-scoring, approach, while PISA represents an attempt to address the problem directly from physical–chemical principles.

Conclusion

Here, we have described the theoretical background of a new approach to the identification of macromolecular complexes in crystals. We have also introduced the new publicly available EBI-MSD service PISA, which implements the method. The software provides a single-button analysis of X-ray-resolved structures, including the assessment of macromolecular interfaces, presence of thermodynamically stable complexes, and their probable dissociation patterns. Structure solution by means of X-ray diffraction on macromolecular crystals remains by far the most common technique in the field as of today, therefore we expect PISA to be of a substantial practical interest to the crystallographic community, and as a tool that allows one to automatically obtain additional and important information from available experimental data.

PISA represents the first, to our knowledge, systematic approach to automatic identification of probable quaternary structures, based on physical–chemical models of macromolecular interactions and chemical thermodynamics. On these grounds, PISA offers a range of features that are not available in more traditional bioinformatic techniques. In particular, the results are delivered in conventional chemical notations, macromolecular interactions are detailed on residue level so that residue substitution effect may be easily modeled and interpreted, symmetry effects are taken into account, and chemical stability of macromolecular complexes is estimated in Gibbs free energy terms along with a suggestion of probable dissociation patterns.

On the training data set, the achieved misclassification error was estimated to be close to ±5 kcal/mol in Gibbs free energy calculations. PISA achieves as many as 90% successful multimeric state classifications on the training data set, which includes about 1.35% of all X-ray entries currently in PDB. As found, the strongest misclassifications in the data set are either due to the difference between experimental and physiological conditions or they result from the use of additional biological functional data for multimeric state assignments. Although the actual absence of gold standard for the latter makes it impossible to rigorously estimate the performance of our method beyond the validated training data set, comparison with partly algorithmic–partly manual annotation in the MSD database[98] shows a reasonable agreement.

In view of the above, we see the primary use of PISA as an aid tool for the analysis and modeling of macromolecular interactions, as well as for the biological unit prediction/assignment, and closely related to that problem of protein functional analysis and annotation. The tool may also be helpful for crystal design, and investigation of residue substitution effects and similar studies.

Finally, we would like to note that, being based on chemical–physical principles, PISA models can be naturally extended in order to account for specific experimental and physiological conditions such as temperature, salinity, pH, and subunit concentration. This possibility was intentionally ignored in the present study primarily in attempt to keep the software product as simple as possible for the user, but also because if additional parameters were introduced, more experimental data would be required to properly calibrate the theoretical models. However, the possibility to take experimental conditions into account appears to be a useful and desired feature, and as such it may make a line of future developments.

Materials and Methods

Protein data bank accession codes

The following PDB entries were used in order to calibrate the protein–protein interaction parameters of Eqs. (1), (9) and (17): 1a19, 1a6q, 1a8o, 1afk, 1ah7, 1ako, 1amj, 1aoh, 1aua, 1aun, 1ayi, 1ayl, 1bc2, 1be0, 1bea, 1bkz, 1bp1, 1bry, 1bwz,1cki, 1ckm, 1ctj, 1dff, 1djx, 1dmr, 1esf, 1fdr, 1feh, 1fsu, 1iae, 1ips, 1kpt, 1kwa, 1lrv, 1mdt, 1mh1, 1mpgk 1np4, 1pda, 1pgs, 1pmi, 1ppo, 1ps1, 1rhs, 1ton, 1xgs, 232l, 2abx, 2acy, 2atj, 2bls, 2hex, 2ihl, 2mbr, 3cms, 1a3c, 1ad3, 1af5, 1afw, 1ajs, 1alk, 1hlr, 1amk, 1aom, 1aor, 1aq6, 1auo, 1bam, 1bif, 1bsr, 1buo, 1cg2, 1chm, 1cmb, 1cp2, 1csh, 1ctt, 1czj, 1daa, 1fip, 1fro, 1gvp, 1hjr, 1hss, 1icw, 1imb, 1isa, 1iso, 1jhg, 1jsg, 1kba, 1kpf, 1lyn, 1mjl, 1mka, 1moq, 1nox, 1nsy, 1oac, 1opy, 1otp, 1pgt, 1pre, 1puc, 1rfb, 1rpo, 1ses, 1slt, 1smn, 1smt, 1sox, 1tox, 1trk, 1tys, 1uby, 1utg, 1wgj, 1xso, 2ccy, 2ilk, 2rsp, 2tct, 2tgi, 3grs, 3pgh, 3sdh, 3ssi, 4kbp, 5csm, 5tmp, 9wga, 1aa0, 1b77, 1ca4, 1cb0, 1cbu, 1ce0, 1cjd, 1dpt, 1dun, 1e2a, 1fgj, 1nif, 1nks, 1ppr, 1qex, 1qlm, 1rla, 2chs, 2pii, 2std, 3cla, 3csu, 3tdt, 4bcl, 1a0l, 1a2z, 1a4e, 1ado, 1az9, 1b25, 1bfd, 1bsm, 1buc, 1bvq, 1cs1, 1cuk, 1dco, 1eta, 1euh, 1ftr, 1gp1, 1gsh, 1ith, 1mpy, 1mxb, 1nhk, 1nhp, 1sml, 1toh, 1uox, 1xva, 2fua, 2izg, 4pga, 5pgm, 1a3g, 1bgv, 1cks, 1dci, 1dxe, 1lcp, 1ndc, 2cev, 2eip, 3gcb, 1ajq, 1b0n, 1ft1, 1h2a, 1hcn, 1hfe, 1ixx, 1luc, 1req, 2frv, 2pka, 4mon, 1apy, 1b7y, 1bou, 1ccw, 1qdl, 1qsh, 2scu, 1eg9, 1mda, and 1mro. This dataset was previously used by Ponstingl *et al.*[12] for calibrating the interaction parameters in PITA software.

Protein–DNA interaction parameters were calibrated using PDB entries 2puc, 2pud, 2pue, 2puf, 2pug, 1vpw, 1qpz, 1zay, 1fok, 1gdt, 1hcr, 1ddn, 1ber, 1cgp, 2cgp, 1run, 1ruo, 1apl, 1yrn, 1fjl, 1hdd, 1au7, 1oct, 2hdd, 3hdd, 9ant, 6pax, 1akh, 1b72, 1mnm, 1ign, 1pdn, 1tc3, 1d3u, 1vol, 1c9b, 2irf, 3hts, 1cf7, 1bc7, 1bc8, 1pue, 1awc, 1zaa, 1aay, 2drp, 1ubd, 1a1g, 1a1h, 1a1i, 1a1j, 1a1k, 1a1l, 2gli, 1glu, 1lat, 1hcq, 2nll, 1by4, 1cit, 1a6y, 1tsr, 1tup, 1d66, 1zme, 1ysa, 1dgc, 2dgc, 1ysa, 1a02, 1an2, 1mdy, 1an4, 1hlo, 1am9, 1a0a, 1crx, 2bop, 1aoi, 1b3t, 1skn, 1qrv, 1ckt, 1ais, 1ytb, 1ytf, 1cdw, 1tgh, 1ecr, 1ihf, 1azp, 1azq, 1bnz, 1bf4, 1bdt, 1bdv, 1par, 1xbr, 1nfk, 1svc, 1a3q, 1vkx, 1ram, 1dct, 1mht, 3mht, 4mht, 5mht, 6mht, 7mht, 8mht, 9mht, 10mh, 1pvi, 2pvi, 3pvi, 1vas, 1rva, 1rvb, 1rvc, 2rve, 4rve, 1rv5, 1bgb, 1bss, 1bua, 1bsu, 1bhm, 3bam, 1eri, 1qps, 1qrh, 1qri, 1cw0, 1tau, 2bdp, 4bdp, 1qsy, 1qss, 2ktq, 3ktq, 4ktq, 1t7p, 1clq, 1dnk, 2dnj, 2bpf, 1bpx, 1bpy, 1bpz, 1zqa, 1zqf, 1zqi, 1zqn, 1zqp, 7ice, 7icg, 7ich, 7ici, 7ick, 7icm, 7icn, 7icp, 7icq, 7icr, 7ics, 7ict, 7icv, 8ica, 8icc, 8icf, 8ici, 8ick, 8icm, 8icn, 8ico, 8icp, 8icq, 8icr, 8ics, 8icu, 8icx, 9ica, 9icf, 9icg, 9ich, 9ick, 9icl, 9icm, 9icn, 9ico, 9icq, 9icr, 9ics, 9ict, 9icu, 9icv, 9icw, 9icx, 9icy, 1ssp, 2ssp, 4skn, 1bnk, 1a73, 1a74, 1cyq, 1ipp, 1bp7, 1a31, 1a35, and 1a36. These entries and corresponding protein–DNA complexes were reviewed by Luscombe *et al.*[83]

All calculations were done using Linux PC with a 1.2 GHz CPU.

Acknowledgements

The authors would like to thank Prof. Joel Janin for reading the manuscript and helpful discussion of vibrational entropy effects. E.K. is supported by the research grant No. 721/B19544 from the Biotechnology and Biological Sciences Research Council (BBSRC) UK.

References

1. Fermi, G., Perutz, M. F., Shaanan, B. & Fourme, R. (1984). The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. *J. Mol. Biol.* **175**, 159–174.
2. Berg, J. M., Tymoczko, J. L. & Stryer, L. (2002). *Biochemistry.* W.H. Freeman and Co, New York.
3. Liu, T. & Chu, B. (2002). Light scattering by proteins. In (Hubbard, A., ed.), pp. 3023–3043, Marcel Dekker Inc., New York.
4. Feigin, L. A. & Svergun, D. I. (1987). *Structure Analysis by Small Angle X-ray and Neutron Scattering.* Plenum Press, New York.
5. Dass, C. (2001). *Principles and Practice of Biological Mass Spectrometry.* John Wiley & Sons, Inc., New York.
6. Svergun, D. I. & Koch, M. H. J. (2002). Advances in structure analysis using small-angle scattering in solution. *Curr. Opin. Struct. Biol.* **12**, 654–660.
7. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242.
8. Cavanagh, J., Fairbrother, W. J., Palmer, A. G., III & Skelton, N. J. (1995). *Protein NMR Spectroscopy, Principles and Practice.* Academic Press, San Diego.
9. Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography.* Academic Press Inc., London.
10. Ponstingl, H., Henrick, K. & Thornton, J. (2000). Discriminating between homodimeric and monomeric proteins in the crystalline state. *Proteins: Struct. Funct. Genet.* **41**, 47–57.
11. Henrick, K. & Thornton, J. (1998). PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* **23**, 358–361.
12. Ponstingl, H., Kabir, T. & Thornton, J. (2003). Automatic inference of protein quaternary structure from crystals. *J. Appl. Crystallogr.* **36**, 1116–1122.

13. Jones, S. & Thornton, J. M. (1996). Principles of protein–protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
14. Argos, P. (1988). An investigation of protein subunit and domain interfaces. *Protein Eng.* **2**, 101–113.
15. Janin, J. & Chothia, C. (1990). The structure of protein–protein recognition sites. *J. Biol. Chem.* **265**, 16027–16030.
16. Jones, S. & Thornton, J. M. (1995). Protein-protein interactions: a review of protein dimer structures. *Prog. Biophys. Mol. Biol.* **63**, 31–65.
17. Miller, S. (1989). The structure of interfaces between subunits ofdimeric and tetrameric proteins. *Protein Eng.* **3**, 77–83.
18. Padlan, E. A. (1990). On the nature of antibody combining sites: unusual structural features that may confer on these sites an enhanced capacity for binding ligands. *Proteins: Struct. Funct. Genet.* **7**, 112–124.
19. Glaser, F., Morris, R. J., Najmanovich, R. J., Laskowski, R. A. & Thornton, J. M. (2006). A method for localizing ligand binding pockets in protein structures. *Proteins: Struct. Funct. Genet.* **62**, 479–488.
20. Gutteridge, A. & Thornton, J. M. (2005). Understanding nature's catalytic toolkit. *Trends Biochem. Sci.* **30**, 622–629.
21. Gutteridge, A., Bartlett, G. J. & Thornton, J. M. (2003). Using a neural network and spatial clustering to predict the location of active sites in enzymes. *J. Mol. Biol.* **330**, 719–734.
22. Tsai, C. J., Lin, S. L., Wolfson, H. & Nussinov, R. (1996). Protein-protein interfaces: architectures and interactions in protein–protein interfaces and in protein cores. Their similarities and differences. *Crit. Rev. Biochem. Mol. Biol.* **31**, 127–152.
23. Lo Conte, L., Chotia, C. & Janin, J. (1999). The atomic structure of protein–protein recognition sites. *J. Mol. Biol.* **285**, 2177–2198.
24. Chakrabarti, P. & Janin, J. (2002). Dissecting protein–protein recognition sites. *Proteins: Struct. Funct. Genet.* **47**, 334–343.
25. Keskin, O., Tsai, C. J., Wolfson, H. & Nussinov, R. (2004). A new, structurally nonredundant, diverse data set of protein interfaces and its implications. *Protein Sci.* **13**, 1043–1055.
26. Ogmen, U., Keskin, O., Aytuna, A. S., Nussinov, R. & Gursoy, A. (2005). PRISM: protein interactions by structural matching. *Nucleic Acids Res.* **33**, W331–W336.
27. Preiβner, R., Goede, A. & Frommel, C. (1998). Dictionary of interfaces in proteins (DIP). Data bank of complementary molecular surface patches. *J. Mol. Biol.* **280**, 535–550.
28. Gong, S., Park, C., Choi, H., Ko, J., Jang, I., Lee, J. *et al.* (2005). A protein domain interaction interface database: InterPare. *BMC Bioinformatics*, **6**, 207–215.
29. Moore, W. J. (1972). *Physical Chemistry.* Prentice-Hall, Inc., Englewood Cliffs, New Jersey.
30. Baker, E. N. & Hubbard, R. E. (1984). Hydrogen bonding in globular proteins. *Prog. Biophys. Mol. Biol.* **44**, 97–179.
31. Janin, J., Miller, S. & Chothia, C. (1988). Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol.* **204**, 155–164.
32. Horton, N. & Lewis, M. (1992). Calculation of the free energy of association for protein complexes. *Protein Sci.* **1**, 169–181.
33. Janin, J. & Rodier, F. (1995). Protein–protein interaction at crystal contacts. *Proteins: Struct. Funct. Genet.* **23**, 580–587.
34. Hermann, R. B. (1972). Theory of hydrophobic bonding. II. The correlation of hydrocarbon solubility in water with solvent cavity surface area. *J. Phys. Chem.* **76**, 2754–2759.
35. Amidon, G. L., Yalkowsky, S. H., Anik, S. T. & Valvani, S. C. (1975). Solubility of nonelectrolytes in polar solvents. V. Estimation of the solubility of aliphatic monofunctional compounds in water using a molecular surface area approach. *J. Phys. Chem.* **79**, 2239–2246.
36. Floris, F. & Tomasi, J. (1989). Evaluation of the dispersion contribution to the solvation energy. A simple computational model in the continuum approximation. *J. Comp. Chem.* **10**, 616–627.
37. Born, M. (1920). Volumes and heats of hydration of ions. *Z. Phys.* **1**, 45–48.
38. You, T. J. & Bashford, D. (1995). Conformation and hydrogen ion titration of proteins: a continuum electrostatic model with conformational flexibility. *Biophys. J.* **69**, 1721–1733.
39. McDonald, I. & Thornton, J. (1994). Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **238**, 777–793.
40. Bahadur, R. P., Chakrabarti, P., Rodier, F. & Janin, J. (2003). Dissecting subunit interfaces in homodimeric proteins. *Proteins: Struct. Funct. Genet.* **53**, 708–719.
41. Xu, D., Tsai, C.-J. & Nussinov, R. (1997). Hydrogen bonds and salt bridges across protein–protein interfaces. *Protein Eng.* **10**, 999–1012.
42. Pace, C., Shirley, B., McNutt, M. & Gajiwala, K. (1996). Forces contributing to the conformational stability of proteins. *FASEB J.* **10**, 75–83.
43. Fersht, A. (1987). The hydrogen bond in molecular recognition. *Trends Biochem. Sci.* **12**, 3214–3219.
44. Horovitz, A., Serrano, L., Avron, B., Bycroft, M. & Fersht, A. (1990). Strength and co-operativity of contributions of surface salt bridges to protein stability. *J. Mol. Biol.* **216**, 1031–1044.
45. Akke, M. & Forsen, S. (1990). Protein stability and electrostatic interactions between solvent exposed charged side chains. *Proteins: Struct. Funct. Genet.* **8**, 23–29.
46. Braxton, S. (1996). Protein engineering for stability. In *Protein Engineering: Principles and Practice* (Cleland, J. & Craik, C., eds), Wiley-Liss, New York; chapt. 11.
47. Betz, S. (1993). Disulphide bonds and the stability of globular proteins. *Protein Sci.* **2**, 1551–1558.
48. Clarke, J. & Fersht, A. (1993). Engineered disulfide bonds as probes of the folding pathway of barnase: increasing the stability of proteins against the rate of denaturation. *Biochemistry*, **32**, 4322–4329.
49. Zauhar, R. J. & Morgan, R. S. (1985). A new method for computing the macromolecular electric potential. *J. Mol. Biol.* **186**, 815–820.
50. Gilson, M. K., Sharp, K. A. & Honig, B. H. (1987). Calculating the electrostatic potential of molecules in solution: method and error assessment. *J. Comp. Chem.* **9**, 327–335.
51. Davis, M. E. & McCammon, J. A. (1989). Solving the finite difference linearized Poisson–Boltzmann equation: a comparison of relaxation and conjugated gradient methods. *J. Comp. Chem.* **10**, 386–391.
52. Davis, M. E. & McCammon, J. A. (1990). Electrostatics in biomolecular structure and dynamics. *Chem. Rev.* **90**, 509–521.

53. Sharp, K. A. & Honig, B. (1990). Electrostatic interactions in macromolecules: theory and applications. *Annu. Rev. Biophys. Biophys. Chem.* **19**, 301–332.
54. Still, W. C., Tempczyk, A., Hawley, R. & Hendrickson, R. (1990). Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112**, 6127–6129.
55. Nicholls, A. & Honig, B. (1991). A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson–Boltzmann equation. *J. Comp. Chem.* **12**, 435–445.
56. Davis, M. E. & McCammon, J. A. (1991). Dielectric boundary smoothing in finite difference solutions of the Poisson equation: an approach to improve accuracy and convergence. *J. Comp. Chem.* **12**, 909–912.
57. Zhou, H.-X. (1993). Boundary element solution of macromolecular electrostatics: interaction energy between two proteins. *Biophys. J.* **65**, 955–963.
58. You, T. J. & Harvey, S. C. (1993). Finite element approach to the electrostatics of macromolecules with arbitrary geometries. *J. Comp. Chem.* **14**, 484–501.
59. Purisima, E. O. & Nilar, S. H. (1995). A simple yet accurate boundary element method for continuum dielectric calculations. *J. Comp. Chem.* **16**, 681–689.
60. Zaloj, V. & Agmon, N. (1998). Diffusion approach to the linear Poisson–Boltzmann equation. *Chem. Phys. Lett.* **284**, 76–86.
61. Bashford, D. & Case, D. A. (2000). Generalized Born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* **51**, 129–152.
62. Chothia, C. (1974). Hydrophobic bonding and accessible surface areas in proteins. *Nature*, **248**, 338–339.
63. Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature*, **319**, 199–203.
64. Ooi, T., Oobatake, M., Nemethy, G. & Scheraga, H. (1987). Accessible surface areas as a measure of thermodynamic parameters of hydration of peptides. *Proc. Natl Acad. Sci. USA*, **84**, 3086–3090.
65. Vila, J., Vasquez, M. & Scheraga, H. (1991). Empirical solvation models can be used to differentiate native from near-native conformations of bovine pancreatic trypsin inhibitors. *Proteins: Struct. Funct. Genet.* **10**, 199–218.
66. Wesson, L. & Eisenberg, D. (1992). Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Sci.* **1**, 227–235.
67. Juffer, A. H., Eisenhaber, F., Hubbard, S. J., Walther, D. & Argos, P. (1995). Comparison of atomic solvation parametric sets: applicability and limitations in protein folding and binding. *Protein Sci.* **4**, 2499–2509.
68. Wang, J. M., Wang, W., Huo, S. H., Lee, M. & Kollman, P. A. (2001). Solvation model based on weighted solvent accessible surface area. *J. Phys. Chem., B*, **105**, 5055–5067.
69. Hou, T., Qiao, X., Zhang, W. & Xu, X. (2002). Empirical aqueoues solvation models based on accessible surface areas with implicit electrostatics. *J. Phys. Chem., B*, **106**, 11295–11304.
70. Jackson, R. M., Gabb, H. A. & Sternberg, M. J. E. (1998). Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. *J. Mol. Biol.* **276**, 265–285.
71. McQuarrie, D. A. (1976). *Statistical Mechanics.* Harper & Row, New York.
72. Page, M. I. & Jencks, W. P. (1971). Entropic contributions to rate accelerations in enzymic and intramolecular reactions and the chelate effect. *Proc. Natl Acad. Sci. USA*, **68**, 1678–1683.
73. Murray, C. W. & Verdonk, M. L. (2002). The consequences of translational and rotational entropy lost by small molecules on binding to proteins. *J. Comput.-Aided Mol. Des.* **16**, 741–753.
74. Mammen, J., Shakhnovich, E. I., Deutch, J. M. & Whitesides, G. M. (1998). Estimating the entropic cost of self-assembly of multiparticle hydrogen-bonded aggregates based on the cyanuric acid melamine lattice. *J. Org. Chem.* **63**, 3821–3830.
75. Finkelstein, A. V. & Janin, J. (1989). The price of lost freedom: entropy of bimolecular complex formation. *Protein Eng.* **3**, 1–3.
76. Kittel, C. (1995). *Introduction to Solid State Physics*, 7th edit. Wiley, New York.
77. Minh, D. D. L., Bui, J. M., Chang, C., Jain, T., Swanson, J. M. J. & McCammon, J. A. (2005). The entropic cost of protein–protein association: a case study on Acetylcholinesterase binding to Fasciculin-2. *Biophys. J.* **89**, L25–L27.
78. Jaynes, E. T. (1992). The Gibbs Paradox. In *Maximum-ntropy and Bayesian Methods* (Erickson, G., Neudorfer, P. & Smith, C. R., eds), pp. 1–22, Kluwer, Dordrecht, Holland.
79. Krissinel, E. & Henrick, K. (2004). Common subgraph isomorphism detection by backtracking search. *Softw. Pract. Exp.* **34**, 591–607.
80. Murakami, M. T., Arruda, E. Z., Melo, P. A., Martinez, A. B., Calil-Elias, S., Tomaz, M. A. *et al.* (2006). Inhibition of myotoxic activity of bothrops asper myotoxin II by the anti-trypanosomal drug suramin. *J. Mol. Biol.* **350**, 416–426.
81. Schuetz, A., Min, J., Antoshenko, T., Chia-Lin Wang, C. L., Allali-Hassani, A., Dong, A. *et al.* (2007). Structural basis of inhibition of the human NAD+-dependent deacetylase SIRT5 by Suramin. *Structure*, **15**, 377–389.
82. Evdokimov, A. G., Anderson, D. E., Routzahn, K. M. & Waugh, D. S. (2001). Unusual molecular architecture of the Yersinia pestis Cytotoxin YopM: a leucine-rich repeat protein with the shortest repeating unit. *J. Mol. Biol.* **312**, 807–821.
83. Luscombe, N. M., Austin, S. E., Berman, H. M. & Thornton, J. M. (2000). An overview of the structures of protein–DNA complexes. *Genome Biol.* **1**, 1–37.
84. Krissinel, E. & Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr., D*, **60**, 2256–2268.
85. Giudice, E., Vrnai, P. & Lavery, R. (2003). Base pair opening within B-DNA: free energy pathways for GC and AT pairs from umbrella sampling simulations. *Nucleic Acids Res.* **31**, 1434–1443.
86. Daune, M. (1999). *Molecular Biophysics: Structure in Motion. Part I.* Oxford University Press, New York.
87. Rossmann, M. G., Mesyanzhinov, V. V., Arisaka, F. & Leiman, P. G. (2004). The bacteriophage T4 DNA injection machine. *Curr. Opin. Struct. Biol.* **14**, 171–180.
88. Rowsell, S., Pauptit, R. A., Tucker, A. D., Melton, R. G., Blow, D. M. & Brick, P. (1997). Crystal structure of carboxypeptidase G_2, a bacterial enzyme with applications in cancer therapy. *Structure*, **5**, 337–347.
89. DiTusa, C. A., Christensen, T., McCall, K. A., Fierke, C. A. & Toone, E. J. (2001). Thermodynamics of metal ion binding. 1. Metal ion binding by wild-type carbonic anhydrase. *Biochemistry*, **40**, 5338–5344.

90. Lubkowski, J. & Wlodawer, A. (1999). Decamers observed in the crystals of bovine pancreatic trypsin inhibitor. *Acta Crystallogr., D*, **55**, 335–337.
91. Wu, H., Lustbader, J. W., Liu, Y., Canfield, R. E. & Hendrickson, W. A. (1994). Structure of human chorionic gonadotropin at 2.6 Å resolution from MAD analysis of the selenomethionyl protein. *Structure*, **2**, 545–558.
92. Lustbader, J. W., Yarmush, D. L., Birken, S., Puett, D. & Canfield, R. E. (1993). The application of chemical studies of human chorionic gonadotropin to visualize its three-dimensional structure. *Endocr. Rev.* **14**, 291–311.
93. Littlefield, O., Korkhin, Y. & Sigler, P. B. (1999). The structural basis for the oriented assembly of a TBP/TFB/promoter complex. *Biochemistry,* **96**, 13668–13673.
94. Guo, F., Gopaul, D. N. & van Duyne, G. D. (1997). Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. *Nature*, **389**, 40–46.
95. Fujinaga, M. & James, M. N. G. (1987). Rat submaxillary gland serine protease, tonin structure solution and refinement at 1.8 Å resolution. *J. Mol. Biol.* **195**, 373–396.
96. Evdokimov, A. G., Anderson, D. E., Routzahn, K. M. & Waugh, D. S. (2001). Unusual molecular architecture of the Yersinia pestis cytotoxin YopM: a leucine-rich repeat protein with the shortest repeating unit. *J. Mol. Biol.* **312**, 807–821.
97. Haramis, A.-P. G. & Perrakis, A. (2006). Selectivity and promiscuity in Eph receptors. *Structure*, **14**, 169–171.
98. Golovin, A., Oldfield, T. J., Tate, J. G., Velankar, S., Barton, G. J., Boutselakis, H. *et al.* (2004). E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.* **32**, D211–D216.
99. Hunt, J. F., van der Vies, S. M., Henry, L. & Deisenhofer, J. (1997). Structural adaptations in the specialized bacteriophage T4 co-chaperonin Gp31 expand the size of the Anfinsen cage. *Cell*, **90**, 361–371.
100. Almog, R., Maley, F., Maley, G. F., MacColl, R. & van Roey, P. (2004). Three-dimensional structure of the R115E mutant of T4-bacteriophage 2′-deoxycytidylate deaminase. *Biochemistry*, **43**, 13715–13723.
101. Kanamaru, S., Leiman, P. G., Kostyuchenko, V. A., Chipman, P. R., Mesyanzhinov, V. V., Arisaka, F. & Rossmann, M. G. (2002). Structure of the cell-puncturing device of bacteriophage T4. *Nature*, **415**, 553–557.
102. Van Raaij, M. J., Schoehn, G., Burda, M. R. & Miller, S. (2001). Crystal structure of a heat and protease-stable part of the bacteriophage T4 short tail fibre. *J. Mol. Biol.* **314**, 1137–1146.
103. Lariviere, L. & Morera, S. (2002). A base-flipping mechanism for the T4 phage beta-glucosyltransferase and identification of a transition-state analog. *J. Mol. Biol.* **324**, 483–490.
104. Leiman, P. G., Kostyuchenko, V. A., Shneider, M. M., Kurochkina, L. P., Mesyanzhinov, V. V. & Rossmann, M. G. (2000). Structure of bacteriophage T4 gene product 11, the interface between the baseplate and short tail fibers. *J. Mol. Biol.* **301**, 975–985.
105. Thomassen, E., Gielen, G., Schutz, M., Schoehn, G., Abrahams, J. P., Miller, S. & van Raaij, M. J. (2003). The structure of the receptor-binding domain of the bacteriophage T4 short tail fibre reveals a knitted trimeric metal-binding fold. *J. Mol. Biol.* **331**, 361–373.
106. Kostyuchenko, V. A., Navruzbekov, G. A., Kurochkina, L. P., Strelkov, S. V., Mesyanzhinov, V. V. & Rossmann, M. G. (1999). The structure of bacteriophage T4 gene product 9: the trigger for tail contraction. *Struct. Fold Des.* **7**, 1213–1222.
107. Moarefi, I., Jeruzalmi, D., Turner, J., O'Donnell, M. & Kuriyan, J. (2000). Crystal structure of the DNA polymerase processivity factor of T4 bacteriophage. *J. Mol. Biol.* **296**, 1215–1223.
108. Papanikolopoulou, K., Teixeira, S., Belrhali, H., Forsyth, V. T., Mitraki, A. & van Raaij, M. J. (2004). Adenovirus fibre shaft sequences fold into the native triple beta-spiral fold when N-terminally fused to the bacteriophage T4 fibritin foldon trimerisation motif. *J. Mol. Biol.* **342**, 219–227.
109. Leiman, P. G., Shneider, M. M., Mesyanzhinov, V. V. & Rossmann, M. G. (2006). Evolution of bacteriophage tails: structure of T4 gene product 10. *J. Mol. Biol.* **358**, 912–921.
110. Leiman, P. G., Shneider, M. M., Kostyuchenko, V. A., Chipman, P. R., Mesyanzhinov, V. V. & Rossmann, M. G. (2003). Structure and location of gene product 8 in the bacteriophage T4 baseplate. *J. Mol. Biol.* **328**, 821–833.
111. Raaijmakers, H., Toro, I., Birkenbihl, R., Kemper, B. & Suck, D. (2001). Conformational flexibility in T4 endonuclease VII revealed by crystallography: implications for substrate binding and cleavage. *J. Mol. Biol.* **308**, 311–323.
112. Raaijmakers, H., Vix, O., Toro, I., Golz, S., Kemper, B. & Suck, D. (1999). X-ray structure of T4 endonuclease VII: a DNA junction resolvase with a novel fold and unusual domain-swapped dimer architecture. *EMBO J.* **18**, 1447–1458.
113. Finnin, M. S., Cicero, M. P., Davies, C., Porter, S. J., White, S. W. & Kreuzer, K. N. (1997). The activation domain of the MotA transcription factor from bacteriophage T4. *EMBO J.* **16**, 1992–2003.
114. Mueser, T. C., Jones, C. E., Nossal, N. G. & Hyde, C. C. (2000). Bacteriophage T4 gene 59 helicase assembly protein binds replication fork DNA. The 1.45 Å resolution crystal structure reveals a novel alpha-helical two-domain fold. *J. Mol. Biol.* **296**, 597–612.
115. Morera, S., Imberty, A., Aschke-Sonnenborn, U., Ruger, W. & Freemont, P. S. (1999). T4 phage beta-glucosyltransferase: substrate binding and proposed catalytic mechanism. *J. Mol. Biol.* **292**, 717–730.
116. He, M. M., Wood, Z. A., Baase, W. A., Xiao, H. & Matthews, B. W. (2004). Alanine-scanning mutagenesis of the beta-sheet region of phage T4 lysozyme suggests that tertiary context has a dominant effect on beta-sheet formation. *Protein Sci.* **10**, 2716–2724.
117. Sickmier, E. A., Kreuzer, K. N. & White, S. W. (2004). The crystal structure of the UvsW helicase from bacteriophage T4. *Structure*, **12**, 583–592.
118. Lariviere, L., Gueguen-Chaignon, V. & Morera, S. (2003). Crystal structures of the T4 phage beta-glucosyltransferase and the D100A mutant in complex with UDP-glucose: glucose binding and identification of the catalytic base for a direct displacement mechanism. *J. Mol. Biol.* **330**, 1077–1086.
119. Morera, S., Lariviere, L., Kurzeck, J., Aschke-Sonnenborn, U., Freemont, P. S., Janin, J. & Ruger, W. (2001). High resolution crystal structures of T4 phage beta-glucosyltransferase: induced fit and effect of substrate and metal binding. *J. Mol. Biol.* **311**, 569–577.

120. Park, H. J., Yang, C., Treff, N., Satterlee, J. D. & Kang, C. (2002). Crystal structures of unligated and CN-ligated Glycera dibranchiata monomer ferric hemoglobin components III and IV. *Proteins: Struct. Funct. Genet.* **49**, 49–60.

121. Li, N., Sickmier, E. A., Zhang, R., Joachimiak, A. & White, S. W. (2002). The MotA transcription factor from bacteriophage T4 contains a novel DNA-binding domain: the 'double wing' motif. *Mol. Microbiol.* **43**, 1079–1088.

Edited by M. Sternberg

(Received 20 November 2006; received in revised form 26 March 2007; accepted 8 May 2007)
Available online 13 May 2007