Inference of macromolecular assemblies from crystalline state.

Evgeny Krissinel* and Kim Henrick

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

Pre-review version of the manuscript accepted in *Journal of Molecular Biology* on 08.05.2007

Abstract

We discuss basic physical-chemical principles underlying 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 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 available at
`http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html`

Keywords: macromolecular assembly, X-ray crystallography, asymmetric unit, biological unit, protein-protein interactions, entropy of macromolecular complexation.

*Corresponding author. E-mail address: keb@ebi.ac.uk

Contents

List of Figures

List of Tables

1 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. Functionality of many, if not most, proteins is not independent 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].

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 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 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 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 neighbouring 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 crystal, are classified into "biologically relevant" and "insignificant" (crystal packing) ones according to a certain scoring system (cf., e.g., Ref. [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 PQS server at MSD-EBI [11], which builds assemblies by progressive addition of suitable chain contacts. Another approach is represented by PITA 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 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 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 detail study of interface properties reported in Ref. [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, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26], but no universal criteria were found. A few databases of protein-protein interactions and interfaces, derived from PDB, have been developed [25, 26, 27, 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 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.

2 Macromolecular complexes in solutions

Complexes differ from molecules in that, typically, their subunits do not make strong 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. Biological units are most often viewed as stable structures. However, because of relatively weak binding of their subunits, complexes, in general, exist in a dynamic equilibrium between different multimeric forms, which equilibrium is subject to subunit concentrations, temperature and solvent properties.

In order to better understand different factors, affecting formation and dissociation of macromolecular complexes, relevant to the problem of their identification in crystallized form, in this Section we discuss thermodynamical basics of macromolecular complexation and give interpretation of complex stability. We further consider principles of macromolecular binding in solutions and role of entropy in complex formation.

2.1 Stability of macromolecular complexes

Consider complex $A = (A_1, A_2 \dots A_n)$ made of subunits A_i. Any subunit may be a complex or a monomeric unit, such as 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}. Two interfaces are considered to be equivalent if they are formed by structurally and chemically equivalent subunits in equal relative positions. Equivalent interfaces are combined into interface equivalence types.

Dissociation of complex A into subunits A_i involves a change in the standard Gibbs free energy, given by the following expression [29]

$$\Delta G^0_{diss} = -\Delta G_{int} - T\Delta S \tag{1}$$

where ΔG_{int} is free energy gain due to bringing subunits A_i together in the complex (binding energy), T is absolute temperature and ΔS represents entropy change upon dissociation. ΔG^0_{diss} depends on many factors, in particular, on the set of subunits $\{A_1, A_2 \dots A_n\}$, which we will call as "dissociation pattern". Since we consider only equilibrium states, dissociation patterns may include only stable subunits. In addition, dissociation patterns should satisfy interface symmetry, which means that if one interface of a particular equivalence type is disengaged upon dissociation, all other interfaces of the same type are also disengaged.

Generally speaking, dissociation may proceed along different patterns if they correspond to close values of ΔG^0_{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 ΔG^0_{diss} is realized. One may hope that this assumption holds in most instances, because existence of different patterns with close values of ΔG^0_{diss} seems to be rather exceptional in real world.

Standard free energy of dissociation ΔG^0_{diss} may be expressed in terms of the equilibrium dissociation constant K_d [29]:

$$\Delta G^0_{diss} = -RT \log K_d = -RT \log \frac{\prod_{i=1}^{n} [A_i^0]}{[A^0]} \tag{2}$$

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. The ability of complex to dissociate at arbitrary concentrations is measured by free energy of dissociation [29]

$$\Delta G_{diss} = \Delta G^0_{diss} + RT \log \frac{\prod_i [A_i]}{[A]} \tag{3}$$

If ΔG_{diss} is positive, then dissociation requires work to be done on the system. At negative ΔG_{diss}, the dissociation proceeds spontaneously. Using Eqs.

(2,3), one can define complex as stable *in the standard state* if $K_d < 1$, or if $\Delta G^0_{diss} > 0$. It is important to realize that this definition does not, in general, imply prevalence of equilibrium complex concentration over that of subunits in normal physiological conditions. As may be found from Eq. (2), assuming that neither of subunits is found in an excess amount, $[A^0] \geq [A^0_i]$ is attained at $\Delta G^0_{diss} \geq -(n-1)RT \log[A^0_i]$. In case of dimers ($n = 2$), at $[A^0_i] = 1$ mM this yields $\Delta G^0_{diss} \geq 4.1$ kcal/mol. As is shown later in the paper, this figure is close to the expected error in the estimated standard free energies, therefore, in many cases the above definition of complex stability will imply prevalence of complex concentration.

2.2 Macromolecular binding in solutions

Consider now the binding term of basic equation (1). There are many factors that contribute into affinity of macromolecules [13, 14, 15, 16, 17, 30, 31, 32, 33]. Generally, the free energy of binding ΔG_{int} originates from the change of solvation energy ΔG_{solv} and immediate (contact-dependent ΔG_{cont} and electrostatic ΔG_{es}) interactions between the subunits upon bringing them together in the complex:

$$\begin{aligned} \Delta G_{int}(A) \quad = \quad & \Delta G_{solv}(A) - \sum_{i=1}^{n} \Delta G_{solv}(A_i) + \\ & \sum_{j>i} \Delta G_{cont}(A_i, A_j) + \sum_{j>i} \Delta G_{es}(A_i, A_j) \end{aligned} \tag{4}$$

Any complex may be viewed as a structure made of subunits A_i connected by interfaces. It is important to see that $\Delta G_{int}(A)$ *does not* reduce to the sum of binding energies of all interfaces in the complex, $\Delta G_{int}(I_{ij})$, found as binding energies of two-subunit complexes $A_{ij} = (A_i, A_j)$. In order to show that, consider terms of Eq. (4) in more details.

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

$$\begin{aligned} \Delta G_{solv}(A) \quad = \quad & \Delta G_{cont}(A, V) - \Delta G_{cont}(C, V) + \\ & \Delta G_{es}(A, V) - \Delta G_{es}(C, V) \end{aligned} \tag{5}$$

where V stands for bulk solvent and all energies represent differences between vacuum and solvent environments.

In Eq. (5), $\Delta G_{cont}(A, V)$ and $\Delta G_{cont}(C, V)$ include any sort of contact interactions, such as hydrogen bonding, between bulk solvent V and macromolecule A or cavity C, respectively. Contact interactions include also the

entropy change arising from full or partial immobilization of solvent molecules due to binding to macromolecular surface. One can reasonably assume that contact interactions are proportional to the cavity surface area σ_A. Electrostatic interactions $\Delta G_{es}(A,V)$ and $\Delta G_{es}(C,V)$ 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, as a result, have a high value of dielectric constant $\epsilon_s \approx 78$. Coulomb interaction in water fades on distance scale of 7 Å (as given by the Onsager length $r_C = e^2/(4\pi\epsilon_0\epsilon_s k_B T)$), which is relatively short comparing to the dimensions of biological macromolecules. Therefore, we may assume that $\Delta G_{es}(C,V)$ depends mainly on the cavity surface area as well. As a result,

$$\Delta G_{solv}(A) \approx \omega\sigma_A + \Delta G_{es}(A,V) \tag{6}$$

This conclusion is supported by experimental measurements of solvation energies of saturated hydrocarbons in water ($\Delta G_{es}(A,V) \approx 0$ [34, 35, 36]), which suggest the value of $\omega \approx 7$ cal/(mol Å^2).

Electrostatic interaction $\Delta G_{es}(A,V)$ is given by the difference of electrostatic energies of macromolecule A in vacuum and solvent environments [37, 38]:

$$\Delta G_{es}(A,V) = \frac{1}{2}\sum_i q_i \sum_j \left(\phi_j^\epsilon(\vec{x}_i) - \phi_j^0(\vec{x}_i)\right) = \frac{1}{2}\sum_i q_i \sum_j \phi_j(\vec{x}_i) \tag{7}$$

where q_i denote atomic charges in positions $\vec{x}_i$, and $\phi_j(\vec{x})$ stands for the difference of electrostatic potentials $\phi_j^\epsilon(\vec{x})$ and $\phi_j^0(\vec{x})$, induced by charge q_j in point $\vec{x}$ in solvent and vacuum environments, respectively. The partial potential $\phi_j(\vec{x})$ may be calculated as Green's function of the Poisson equation:

$$\nabla\epsilon_0\left(\epsilon(\vec{x}) - 1\right)\nabla\phi_j(\vec{x}) = -q_j\delta(\vec{x} - \vec{x_j}) \tag{8}$$

The difference between $\phi_j^\epsilon(\vec{x})$ and $\phi_j^0(\vec{x})$ is due to solvent polarisation induced by the Coulomb interaction of solvent molecules with atomic charges q_i. This is formally accounted for in Eq. (8) by coordinate-dependent dielectric constant $\epsilon(\vec{x})$: $\epsilon(\vec{x}) = \epsilon_M \approx 4$ inside the macromolecule, $\epsilon(\vec{x}) = \epsilon_s \approx 78$ in bulk solvent and $\epsilon(\vec{x}) = 1$ in vacuum. The low value of ϵ_M suggests a relatively long range of electrostatic interactions inside the macromolecule ($r_C \approx 140$Å). Therefore, $\Delta G_{es}(A,V)$ contains contributions from all atomic charges and cannot be approximated by surface interactions. This fact means that solvation energy is not a sole function of macromolecular surface, which leads us to conclusion that a) the binding energy of complex A, $\Delta G_{int}(A)$, is not a sum of interface binding energies $\Delta G_{int}(I_{ij})$, and b) the interface binding energy $\Delta G_{int}(I_{ij})$ is not a function of interface properties only.

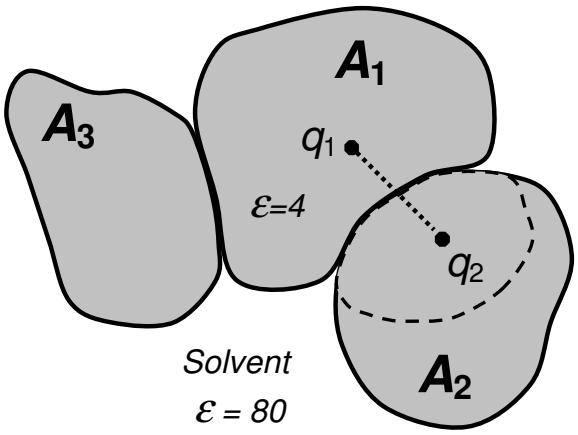

Figure 1: Electrostatic interaction of charges q_1 and q_2 depends on the presence of subunit A_3 due to the difference between dielectric constants ϵ inside the subunits and in the solvent. Therefore, binding energy of complex $A = (A_1, A_2, A_3)$, $\Delta G_{int}(A)$, is not a sum of binding energies of isolated interfaces $\Delta G_{int}(I_{12})$ and $\Delta G_{int}(I_{13})$. For the same reason, interaction between q_1 and q_2 also depends on the size and shape of the subunits, therefore interface binding energy is not a sole function of interface properties. See discussion in the text.

Indeed, consider a model complex in Figure 1. Electrostatic interaction between charges q_1 and q_2, belonging to subunits A_1 and A_2 is given by the following expression:

$$V_{12} = \left(q_1\phi_2(\vec{x}_1) + q_2\phi_1(\vec{x}_2)\right)/2 \tag{9}$$

where partial potentials $\phi_1(\vec{x})$ and $\phi_2(\vec{x})$ satisfy Eq. (8) with $\epsilon(\vec{x})$ corresponding to the geometry of the whole macromolecular complex. Because of the latter, V_{12} depends on the presence of subunit A_3; one may show that V_{12} is larger in the presence of A_3. Similar considerations may be applied to the comparison of identical interfaces formed by otherwise different macromolecules. If part of subunit A_2 in Fig. 1 is removed, as shown by dashed line, electrostatic interaction between charges q_1 and q_2 will decrease. This will change the binding energy of subunits A_1 and A_2, although all properties of interface I_{12}, such as area, shape, geometrical complementarity, chemical composition etc. remain unchanged.

Consider now the contact interaction between subunits $\Delta G_{cont}(A_i, A_j)$ (cf. Eq. (4)). This interaction is mostly due to the formation of hydrogen bonds, salt bridges and disulphide bonds across the interface between A_i and A_j. It may be approximated by the following expression:

$$\Delta G_{cont}(A_i, A_j) = N_{hb}^{ij} E_{hb}^0 + N_{sb}^{ij} E_{sb} + N_{db}^{ij} E_{db} \tag{10}$$

where N_{hb}^{ij}, N_{sb}^{ij}, N_{db}^{ij} stand for the number of hydrogen bonds, salt bridges and disulphide bonds between subunits A_i and A_j, and E_{hb}^0, E_{sb}, E_{db} are the average free energy gains per corresponding bond. Experimental and theoretical studies suggest that $E_{hb}^0 \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 $\Delta G_{int}(A)$. 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. (4) by the sum of terms $\Delta G_{cont}(A_i, V)$ substituted from Eq. (5). 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_{hb} \approx$0.6-1.5 kcal/mol per bond only [42, 43]. Limited experimental data on the stabilisation effect of salt bridges suggest that free energy contribution of a salt bridge is close to that of a hydrogen bond $E_{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 disulphide bond may contribute up to 2-8 kcal/mol [46, 47, 48], at a yet lower occurrence. Due to infrequent occurence of disulphide bonds, they are neglected in our practical calculations, however we leave them formally in final expressions.

It follows from the above that binding energy of a complex may be represented as

$$\Delta G_{int}(A) = \sum_{j>i} \left(-2\omega\Delta\sigma_{ij} + N_{hb}^{ij}E_{hb}^0 + N_{sb}^{ij}E_{sb} + N_{db}^{ij}E_{db}\right) + \Delta G_{es}^* \qquad (11)$$

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}. ΔG_{es}^* accounts for all interactions that cannot be attributed to isolated interfaces:

$$\Delta G_{es}^* = \Delta G_{es}(A, V) - \sum_i \Delta G_{es}(A_i, V) + \sum_{j>i} \Delta G_{es}(A_i, A_j) \qquad (12)$$

As follows from Eqs. (11,12), representation of binding energy $\Delta G_{int}(A)$ as a sum of interface binding energies $\Delta G_{int}(I_{ij})$ 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 have been developed for the solution of Poisson equation (8), which are applicable if atomic charges q_i are known [38, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61]. However, the latter is a genuine problem, because charge distribution in a macromolecule depends on the molecule's environment. Most proteins

represent charged particles due to the presence of charged residues, such as Arg, in their composition. However, the particular charge exerted on the molecule depends on the 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 expression similar to the first term in Eq. (11):

$$\Delta G_{int}(A) \approx \sum_{j>i} \left(\sum_k \omega_k \Delta\sigma_{ij}^k + N_{hb}^{ij} E_{hb}^0 + N_{sb}^{ij} E_{sb} + N_{db}^{ij} E_{db} \right) \tag{13}$$

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 atomic solvation parameter (ASP) for kth atom type. The concept of ASPs has been used in many studies [62, 63, 64, 65, 66, 67, 68, 69]). As may be concluded from the above, ASPs do not have a real physical meaning and should be viewed as empirical parameters allowing for better approximation of electrostatic term ΔG_{es}^* in Eq. (11). Precision of solvation energy estimates with different ASP models was found to be within a few kcal/mol [63, 67, 69], which is similar to that obtainable with electrostatic models [54, 70].

2.3 Entropy of complex dissociation

Entropy S of a thermodynamic system is a logarithmic measure of the number of energy states, or phase volume Φ, that is accessible to the system [71]:

$$S = R \log \Phi \tag{14}$$

Phase evolution of complex A may be represented as a correlated motion of subunits A_i, therefore part of phase volume $\Delta\Phi$, associated with relative motion of the subunits, is not accessible to the system in bound state. Upon dissociation, the system regains phase volume $\Delta\Phi$, which results in entropy change

$$\Delta S = \sum_i S(A_i) - S(A) \tag{15}$$

The corresponding effect on Gibbs free energy amounts to $-T\Delta S$ (cf. Eq. (1)), which may be viewed as entropy driving force of dissociation. If $T\Delta S$ outweights the binding energy, $-\Delta G_{int}(A)$, the system moves toward dissociated state. A simplified physical interpretation of entropy driving force

corresponds to the regain of energy of those degrees of freedom that got lost due to complex formation. Upon binding, their energy is dissipated by thermal bath of the solvent and then is returned to the subunits in the course of dissociation through encounters with the solvent molecules.

Calculation of absolute entropy S is a challenging task, which is not fully solved as of today [71, 72, 73, 74, 75]). Entropy of subunit A may be represented as

$$S(A) = S_{trans}(A) + S_{rot}(A) + S_{vib}(A) + S_{surf}(A) \tag{16}$$

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, 72, 73, 74]:

$$S_{trans}(A) = R \log \left[\left(\frac{2\pi m(A) k_B T}{h^2} \right)^{3/2} \left(v e^{5/2} \right) \right] \tag{17}$$

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 into Sackur-Tetrode equation in Ref. [74] in attempt to account for the reduction of phase space in condensed media. In ideal gas, when volume of molecules is negligibly small comparing to the total volume of the system, v is inversely proportional to the concentration of molecules, $v \approx 1/[A]$. In condensed media, the volume of molecules itself can no longer be neglected, and free volume becomes a complex function of solvent and solute structures and concentration $[A]$. As shown in Ref. [74], setting $v = 1/[A]$ yields values of S as precise as 0.5% of experimental values for real gases, while for liquids it results in $30 - 50\%$ overestimation. Using an appropriate function for v allows one to bring the calculated values of S to within 2% of the observed ones in liquids as well [74], however the parameter remains largely of empiric nature. As will be shown later in the paper, we do not use any explicit value for the free volume v, and leave it hidden in a more general empirical parameter chosen to describe a set of experimentally verified macromolecular assemblies.

Rotational entropy term $S_{rot}(A)$ can be calculated as logarithm of phase volume Φ_r corresponding to the rotational motion of A. The volume is proportional to the number of elementary h^3 cells, attainable by rotational motion, divided by symmetry number $\gamma(A)$, equal to the number of different

rotations that superpose A over itself. The resulting expression reads [71, 73]

$$S_{rot}(A) = R \log \left[\frac{\sqrt{\pi}}{\gamma(A)} \left(\frac{8\pi^2 k_B T e}{h^2} \right)^{3/2} \sqrt{J_1(A) J_2(A) J_3(A)} \right] \qquad (18)$$

where J_1, J_2 and J_3 stand for the principle moments of inertia of unit A. Comparison of Eqs. (17) and (18) indicates that $1/\gamma$ plays the role of free volume in rotational motion, which depends neither on the concentration $[A]$ nor structure of solvent molecules. This is a probable reason for that rotational entropies in liquids differ by only 2% of gas phase values [74], and Eq. (18) is found to be a good approximation in both gases and liquids. Note that symmetry number is omitted in the expression for rotational entropy given in Ref. [74]. We stress that γ is a very important parameter, significance of which in the context of macromolecular complexation will be shown below.

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

$$S_{vib} = R \sum_k \left[\frac{\beta_k}{e^{\beta_k} - 1} - \log\left(1 - e^{-\beta_k}\right) \right], \quad \beta_k = \frac{hc\lambda_k}{k_B T} \qquad (19)$$

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 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 $S_{trans} + S_{rot}$ may be transformed into S_{vib}, however recent MD studies showed a considerably smaller effect [77]. Calculation of vibration spectra $\{\lambda_k\}$ for large asymmetric structures, like proteins, is 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. (16), $S_{surf}(A)$, 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:

$$S_{surf}(A) = F\sigma_A \tag{20}$$

Substituting Eqs. (17,18) and (20) into Eq. (16), and neglecting the vibrational entropy term S_{vib}, obtain

$$S(A) = C + \frac{3}{2}R\log\left(m(A)\right) + \frac{1}{2}R\log\left(\frac{J_1(A)J_2(A)J_3(A)}{\gamma^2(A)}\right) + F\sigma_A \tag{21}$$

where we introduce a constant parameter C, which we consider as empirical one, even though analytical expression for it follows directly from Eqs. (17,18). There are several reasons to leave C as a free parameter in further analysis. Firstly, it depends on the poorly defined free volume v (cf. Eq. (17)), which is essentially an empirical parameter itself. Secondly, C depends on the concentration $[A]$, which we would like to exclude from the consideration in attempt to keep the software as simple for use as possible. Analytical form of function $C([A])$ is not well-defined because of uncertainty in free volume v, except for the fact that $C([A]) \approx a - b\log([A])$ in the limit of low $[A]$, when the molecule's own volume may be neglected. As was pointed out in Section 2.1, our definition of 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, concentration correction may be done using Eq. (3), where second summand represents classical (zero own volume) dependence of entropy change $T\Delta S$ on the concentrations (if functions $v([A])$ were known, one could have more accurate correction by replacing $[A]$ with $1/v([A])$ in Eq. (3)). 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 adjustable parameter may hopefully compensate, at least partially, for these factors.

Using Eq. (21) in Eq. (15), obtain

$$\Delta S \;=\; (n-1)C \;+\; \Delta S^{\star}_{trans} \;+\; \Delta S^{\star}_{rot} \;+\; 2F\sum_{j>i}\Delta\sigma_{ij} \tag{22}$$

where

$$\Delta S^{\star}_{trans} \;=\; \frac{3}{2}R\log\left(\frac{\prod_i m(A_i)}{m(A)}\right) \tag{23}$$

$$\Delta S^{\star}_{rot} \;=\; \frac{1}{2}R\log\left(\frac{\prod_{i;k} J_k(A_i)\gamma^2(A)}{\prod_k J_k(A)\prod_i \gamma^2(A_i)}\right) \tag{24}$$

Expression (22) 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

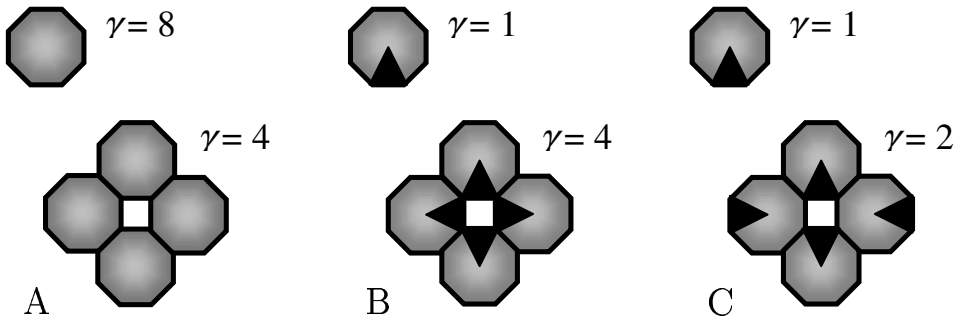

Figure 2: Symmetry effect on the formation of macromolecular complexes. For simplicity, 2D-space is implied in this Figure. A) formation of complex with symmetry number $\gamma(A) = 4$ from 4 subunits with symmetry numbers $\gamma(A_i) = 8$ each is assisted by negative entropy change ΔS^{γ}_{rot} (cf. Eq. (25)) B) formation of the same complex from subunits with symmetry numbers $\gamma(A_i) = 1$ each is inhibited by positive ΔS^{γ}_{rot} C) example of a less-symmetric complex, which may be preferable to one in Fig. 2B, see discussion in the text. Note that, due to symmetry, dissociation patterns of all three complexes ought to be the same (4 original subunits).

interfaces. This stresses the earlier conclusion, made in section 2.2, that complex stability cannot be inferred only from interface analysis.

According to the second law of thermodynamics, an isolated thermodynamic system always tends to maximize its entropy, or phase volume Φ attained [71]. Maximal Φ corresponds to the most disordered (dissolved) state possible, which makes entropy the driving force of dissociation. It is interesting to see that entropy also drives macromolecular complexes to less-symmetric states.

Eq. (24) may be split into sum of two terms depending on the principal moments of inertia, ΔS_J, and symmetry numbers, ΔS_γ:

$$\begin{aligned} \Delta S^{\star}_{rot} &= \Delta S_J + \Delta S_\gamma \\ &= \frac{1}{2} R \log\left(\frac{\prod_{i;k} J_k(A_i)}{\prod_k J_k(A)}\right) + R \log\left(\frac{\gamma(A)}{\prod_i \gamma(A_i)}\right) \end{aligned} \qquad (25)$$

Consider now formation of a complex with symmetry number $\gamma(A) > 1$ from subunits with higher, then complex's, symmetry numbers $\gamma(A_i) \geq \gamma(A)$ (see example in Figure 2A). For such system, $\Delta S_\gamma < 0$ makes positive contribution into the free energy of dissociation ΔG^0_{diss} (cf. Eq. (1)), and therefore complexation may be viewed as symmetry-assisted (and, correspondingly, dissociation is symmetry-inhibited). If the same complex is made of asymmetric subunits with $\gamma(A_i) = 1$ (as in Figure 2B), the complexation is symmetry-

inhibited (and dissociation is symmetry-assisted) due to negative contribution of $\Delta S_\gamma > 0$ into ΔG^0_{diss}. From this, we can conclude that entropy favours lower symmetry states in thermodynamic systems. Indeed, subject to their structural properties, asymmetric subunits may be packed into structurally equivalent complexes with different symmetries (compare Figs. 2B and C). Provided that these complexes have identical dissociation patterns, very close binding energies ΔG_{int} and moments of inertia, the one with lowest symmetry number is most preferable: as may be seen from Eqs. (1,22,25), ΔG^0_{diss} increases with decreasing $\gamma(A)$. In practice, however, ΔG_{int} usually reaches its minimum in the most symmetric packings and overweights the entropy drive to lower symmetries. Also, dissociation pattern usually depends on the packing. As a result, homogeneous complexes are rarely asymmetric.

The symmetry effect on entropy of dissociation has analogy with the celebrated Gibbs paradox [78]: distinguishing between different subunit orientations in Figs. 2B and C, on comparison with indistinguishable orientations in Fig. 2A, changes ΔS similarly to what happens at mixing of like and unlike types of gases. Just like in the classical case, this change does not depend on the nature of subunits, on their interactions and on what makes them asymmetric. The essence of the paradox is that, generally speaking, symmetry numbers γ and, correspondingly, the value of ΔS, depend on what we know about the subunits. Suppose that today our system looks like the one in Fig. 2A with stable complex made of 4 identical subunits. Following discussion in Ref. [78], imagine now that next century brings a discovery of two types of carbon, which makes the subunits asymmetric as in Figs. 2B,C. Then we would have to recalculate the entropy term in Eq. (1) and may well find that the complex should be unstable.

The "paradox" is resolved in the same way as the classical one. As pointed out in Ref. [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. Thus, a mere difference between isotopes of carbon is most probably not enough for decreasing a subunit's symmetry number, unless it results in a drastic change of surface properties or moments of inertia etc., or, e.g., if we deal with magnetic isotopes in (very) strong magnetic fields. In the next Section we list empirical criteria for the identification of similarity and symmetry numbers of subunits and complexes, which are used in our software application. One

can read more details about Gibbs paradox and translate them onto subject in question from a beautiful discussion in Ref. [78].

3 Identification of macromolecular assemblies in crystals

Theoretical analysis, presented in Section 2, can answer the question whether a particular macromolecular assembly is stable, i.e. 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.

3.1 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, $f = [A^0]/[A_i^0] = [A_i^0]^{n-1}/K_d$, only increases with 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 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 crystal may be obtained by repeated translations from one original unit.

This means that any two parallel units, found at a given relative position in crystal, 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, i.e. 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 crystal by enumeration of 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., Ref. [79]). Starting from 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 of 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. E.g., if interfaces I_{12} and I_{23} are engaged in trimer (A_1, A_2, A_3), interaface 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} can not become positive on any of 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 $\Delta G^r_{int} + T\Delta S^r > 0$ 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 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 ΔG_{int}^{r+1} cannot be less than ΔG_{int}^r plus sum of binding energies of all hydrophobic interfaces that can still be engaged on subsequent recursion levels. Hence, the termination condition is

$$\Delta G_{int}^r + T\Delta S^r + \sum_{jk} \min\left(\Delta G_{int}(I_{jk}), 0\right) \quad \geq \quad 0 \tag{26}$$

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 $\Delta G_{int}(I_{jk})$. Having little effect in cases with less than 10 interface types, condition (26) becomes vital for performance if N_I is greater than 20.

3.2 Dissociation pattern

In order to infer on chemical stability of an assembly, one needs to calculate the Gibbs free energy of dissociation ΔG_{diss}^0, using Eqs. (1,13,22). This is possible only if dissociation pattern $\{A_i\}$ is known. As was discussed in Section 2.1, we assume that assemblies dissociate into *stable* subunits along a pattern with minimal ΔG_{diss}^0. In addition, from symmetry considerations, dissociation patern cannot contain contain an engaged interface, if an equivalent interface is disengaged anywhere else in crystal. Dissociation patterns may be found by a backtracking scheme similar to one described in Section 3.1 for assembly search. Represent 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. 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_{diss}^0 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.

3.3 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 2FZQ [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 presence of calcium ions [82]. In general, it is protein-ligand interactions that underly 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 2 of them in the asymmetric unit. If a ligand interfaces with more than one macromolecular unit, the one with 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 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 ΔG^0_{diss}) were different in almost every case, with ligands detaching from a more stable macromolecular assembly when fixing was not used.

3.4 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 can not 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 functional role of macromolecular assembly is not directly connected with dissociation, 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. As follows from the discussion in the beginning of Section 3.1, 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 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 ΔG^0_{diss} take preference over those with lower ΔG^0_{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 thus sorted list of potential answers shall be meant, unless otherwise specified.

3.5 Implementation and calibration of parameters

Model equations for standard Gibbs free energy ΔG^0_{diss}, derived in Section 2 (Eqs. (1,13,22)), include undefined values of atomic solvation parameters ω_k, hydrogen bond, salt bridge and disulphide bond contributions E_{hb}, E_{sb}, E_{db}, respectively, as well as entropy parameters C and F. As discussed in Section 2, no rigorous theoretical expressions may be suggested in order to estimate these parameters, therefore we choose them such as to achieve best agreement with available experimental data on macromolecular complexes.

For the calibration of model parameters, we used previously published dataset of 218 protein structures with experimentally verified multimeric states [12]. Atomic ASPs were chosen according to the scheme suggested in Ref. [63], which classes atoms into 5 types: carbon (C), sulphur (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 Refs. [63, 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 dataset. 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 dataset as possible:

$$\begin{cases} \Delta G^0_{diss} > 0 & \text{for correct assemblies} \\ \Delta G^0_{diss} \leq 0 & \text{for all other assemblies of same or bigger size} \end{cases} \tag{27}$$

Obtained parameters are then used for assembly calculations in the next iteration, and the process stops when the number of dataset 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 in Ref. [83]. The nucleic acid atoms were classed into 8 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 carbons (C_{na}). Only nucleic acid ASPs were varied on this step, with protein ASPs and other parameters taken as resulted from 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 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 algorithm SSM [84]. Two monomeric units were considered as equivalent if structure alignment yields the following values of SSM's quality score Q and sequence identity SI:

$$Q = \frac{N^2_{align}}{(1 + (RMSD/3)^2)N_1N_2)} \geq 0.65 \qquad SI = \frac{N_{ident}}{N_{align}} \geq 0.9 \tag{28}$$

where N_1, N_2 stand for the number of residues in two structures, $RMSD$ is r.m.s.d. between aligned C_α's in best structure superposition, 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 PISA (Protein Interfaces, Surfaces and Assemblies) and made available for public use as an interactive web-server found at `http://www.ebi.ac.uk/msd-srv/prot_int/pistart.html`. 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 seconds, most difficult cases may take up to 20 minutes. The server also provides a detail description of interfaces, structures and assemblies, their visualisation and database search tools.

4 Results and discussion

Table 1 shows empirical parameters of theoretical models for binding energy ΔG_{int} (Eq. (13)) and entropy of dissociation ΔS (Eq. (22)), obtained by fitting procedure described in Section 3.5. As found, the system of inequalities (27) remains underfit for the datasets used, which means that maximal number of satisfied inequalities is achieved by many different sets of fitted parameters. This implies that the datasets 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 aminoacid 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

Aminocid ASPs, cal/(mol Å^2)

		A	B
C	16	16±2	4±3
S	41	21±10	-17±22
O/N	-11	-6±4	-113±14
N^+	-37	-50±9	-169±31
O^-	-17	-24±10	-166±31

Entropy parameters

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

Nucleic acid ASPs, cal/(mol Å^2)

P	-23
O_P	53
O_C	57
O_{na}	23
N_{aro}	0
N_{ami}	-19
C_{ali}	14
C_{na}	-33

Table 1: Empirical parameters entering Eqs. (13,22), obtained by fitting the multimeric states found in the benchmark set of 218 protein complexes from Ref. [12] (table on the left) and 212 protein–DNA complexes from Ref. [83] (table on the right). Value of $T \cdot C$ assumes that mass is measured in a.u. and distance - in Å. Column A shows ASPs obtained in Ref. [63], and column B shows those from Ref. [66].

in columns A and B in Table 1, obtained in Ref. [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 Section 2.2, or it may indicate that 5 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 values of fitted parameters, as long as they satisfy the maximum number of inequalities (27).

As seen from Table 1, energy contributions of hydrogen bonds and salt bridges appear somewhat lower than the estimates found in the literature (cf. Section 2.2), 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 10^3Å^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 molecular dynamics studies (e.g., Ref. [85] suggests 4-7 kcal/mol), but agrees reasonably well with other data (1.6 kcal/mol found in Ref. [86]). The figure

	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%

Table 2: Assembly classification obtained for the benchmark set of 218 PDB entries, representing protein complexes, from Ref. [12]. 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.

of $\sim$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 lowest value of 84% obtained for tetramers. Tetramers have been found as less predictable class also in Ref. [12]. Comparison with benchmark results in Ref. [12] shows an improvement of 5-8% in the classification success rate. On comparison, the PQS server at MSD-EBI [11] achieves 78% correct answers on the same dataset. However, one should take into account

	2mr	3mr	4mr	5mr	6mr	10mr	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%

Table 3: Assembly classification obtained for the benchmark set of 212 PDB entries, representing protein-DNA complexes, reviewed in Ref. [83]. The rows give counts of multimeric states obtained for assemblies annotated as dimeric, trimeric, tetrameric, pentameric, hexameric and decameric in the set.

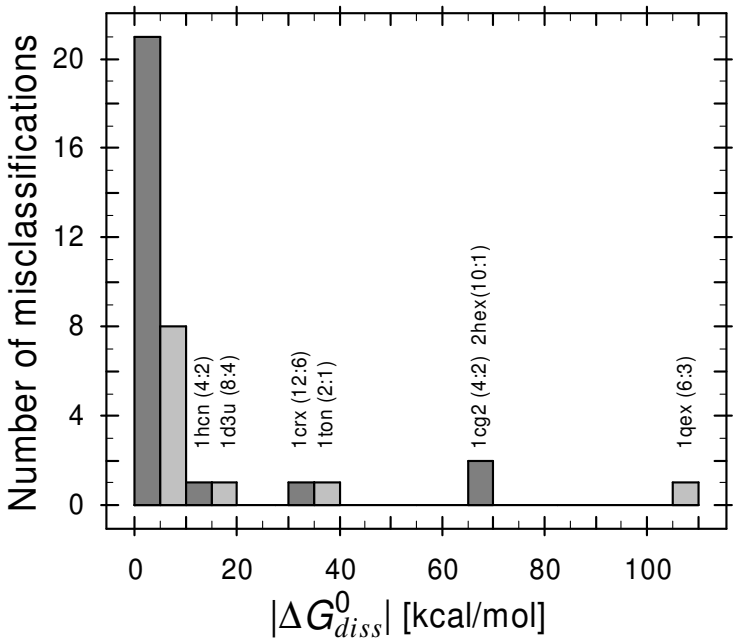

Figure 3: Distribution of misclassified biological units in Tables 2 and 3 over 5 kcal/mol intervals of free energy of dissociation ΔG^0_{diss}. In case of misclassification into a higher multimeric state, the classification error in ΔG^0_{diss} was defined as free energy of dissociation of that state. 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) denote the misclassified (N) and correct (M) multimeric states.

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 dataset.

Classification of protein-DNA complexes shows a somewhat higher success rate (cf. Table 3), with 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 3 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

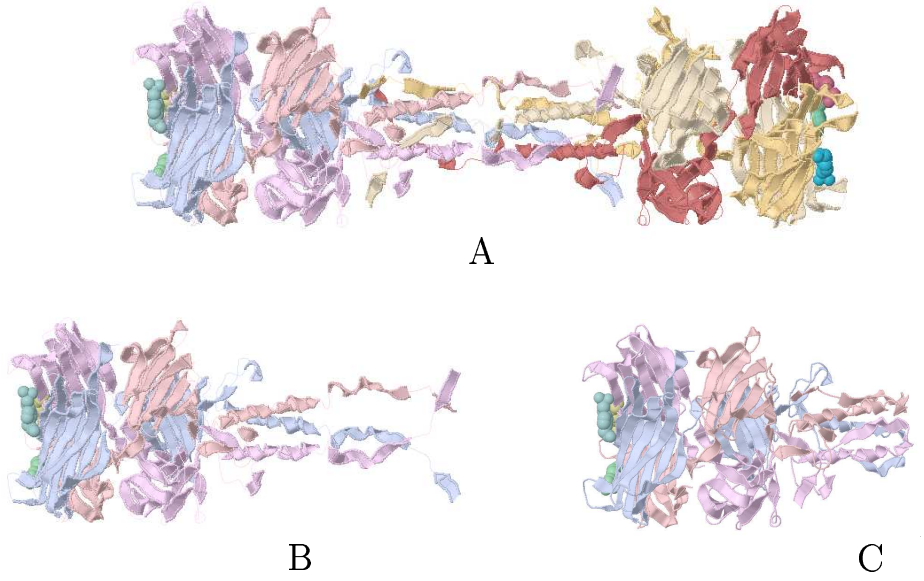

Figure 4: Homo-hexamer predicted for PDB entry 1QEX (A), the corresponding homo-trimeric biological unit (B) identified in Ref. [87], and correct homo-trimer predicted for PDB entry 1S2E (C), representing an alternative fit of the same experimental data. See discussion in the text. The images were obtained using the Jmol software from `http://www.jmol.org`.

the fitting procedure, misclassifications within ±5-10 kcal/mol should not be rated as unexpectable. However, as Figure 3 shows, there are few cases where error in ΔG^0_{diss} is considerably higher than could be reasonably expected even for relatively approximate models 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 4A) dissociates into almost equally stable homotrimers (Figure 4B, $\Delta G^0_{diss} \approx 90$ kcal/mol), which were identified as biological units in Ref. [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 surprizing 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

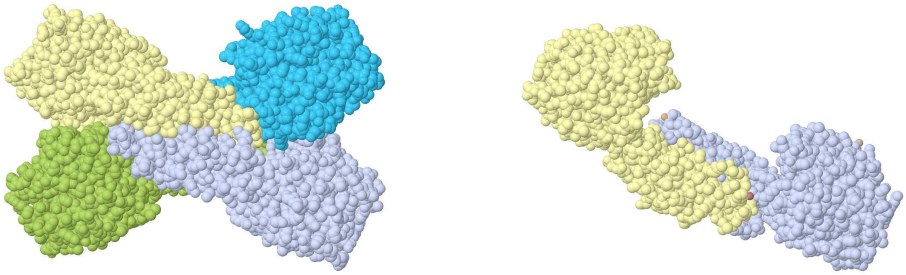

Figure 5: Homo-tetramer predicted for PDB entry 1CG2 (A), and the corresponding homo-dimeric biological unit identified in Ref. [88]. See discussion in the text. The images were obtained using the Jmol software from `http://www.jmol.org`.

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. We therefore conclude that hexameric complex in Figure 4A 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.

Next to 1QEX in free energy classification error, PDB entry 1CG2 was predicted to be a homotetramer shown in Figure 5A instead of a dimer as in Figure 5B. The tetramer is predicted to dissociate into the dimers at $\Delta G^0_{diss} \approx 66$ kcal/mol. This high dissociation barrier is mostly due to the presence of four Zn^{+2} ions 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 Ref. [89]. No Zn binding is involved into 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 neighbours 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 aminoacid composition on the surface may have a pronounced effect on binding properties of proteins and, as a result, cause difference in crystal packing. Indeed, crystal packing of all structural neighbours 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.

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

Table 4: Closest structural neighbours to PDB entry 1CG2 and their assembly classifications, obtained by PISA. Q-score is a measure of structural similarity (cf. Eq. (28), [84]), sequence identity (Seq.Id) was calculated from structure alignment. The neighbours have been identified by structure search facility in PISA software based on SSM algorithm [84]. See discussion in the text.

All structural neighbours 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 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 Ref. [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.

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 3, except for PDB entries 1QEX, 1D3U, 1CRX and 1TON are due to this reason. 2HEX is predicted to be a strong homodecamer, dissociating into 5 homodimers at $\Delta G^0_{diss} \approx 65$ kcal/mol as shown in Figures 6A,B. This decameric structure was noted in Ref. [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 7) dissociates, at $\Delta G^0_{diss} \approx 15$ kcal/mol, into identical heterodimers, which were found to be biological units in Ref. [91]. The dimers appear to be somewhat more stable than the tetramer ($\Delta G^0_{diss} \approx 20$ kcal/mol). Formation of the tetramer may possibly be assisted by the binding sites formed by residues 38-57 (shown in red in Fig. 7), which were identified by antibody blocking in Ref. [92].

Results for PDB entry 1D3U (cf. Figure 8) exemplify a situation that

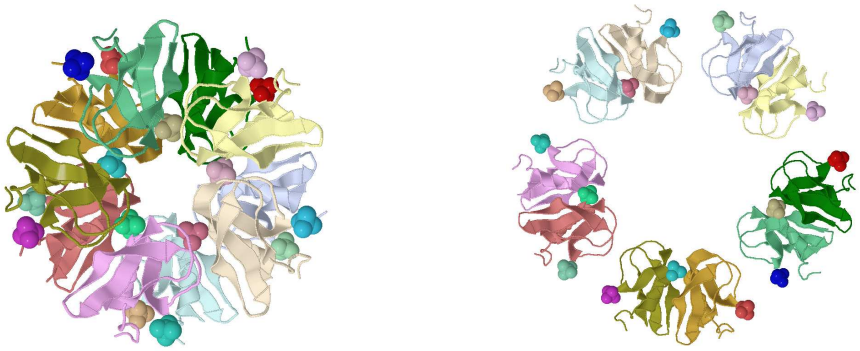

Figure 6: Homo-decamer predicted for PDB entry 2HEX (A), and its dissociation into 5 homo-dimers suggested by PISA analysis (B). The space-fill mode shows SO_4 molecules. See discussion in the text. The images were obtained using the Jmol software from `http://www.jmol.org`.

is specific to protein-DNA complexes. The predicted octamer dissociates into two identical tetramers, identified as biological units in Ref. [93], at $\Delta G^0_{diss} \approx 20$ 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 Fig. 8. In this case, PISA prediction fails because, strictly speaking, neither the crystallized octamer nor tetramer represent 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, can not occur in natural conditions. As estimated by PISA, the interface between two contacting helices of the left- and right-side tetramers in Fig. 8 shows a relatively high hy-

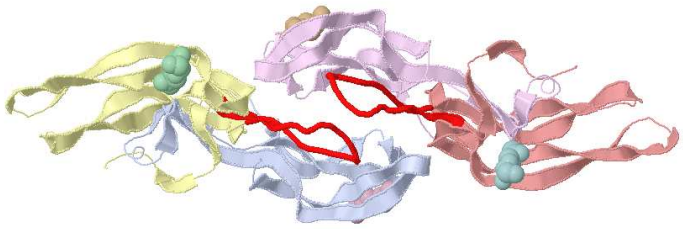

Figure 7: Hetero-tetramer predicted for PDB entry 1HCN, made of two identical hetero-dimers found to be biological units in Ref. [91]. The space-fill mode shows NAG molecules, red color denotes residues identified by antibody blocking (cf. Ref. [92]). See discussion in the text. The image was obtained using the Jmol software from `http://www.jmol.org`.

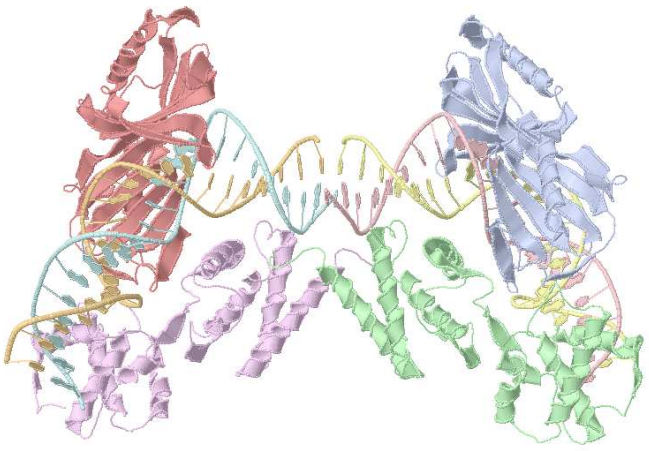

Figure 8: Hetero-octamer predicted for PDB entry 1D3U. The structure dissociates into two identical hetero-tetramers (the biological units, cf. Ref. [93]), making right- and left-hand parts of the shown complex. See discussion in the text. The image was obtained using the Jmol software from `http://www.jmol.org`.

drophobic effect of $\Delta G_{sol} \approx -7.4$ kcal/mol and forms 18 hydrogen bonds and 8 salt bridges, which add further ≈ -9 kcal/mol to the interface binding energy ΔG_{int}. In addition, as may also be seen in Fig. 8, the DNA strands of the tetramers make a cross-pairing, which involves 3 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 9. 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 Fig. 9, at $\Delta G^0_{diss} \approx 28$ kcal/mol, although, from symmetry considerations, one would expect dissociation into trimeric state. A detail examination reveals that dissociation into hexamers is due to asymmetry caused by the presence of modified O-phosphotyrosine residues in 2 out of 4 protein chains, shown in space-fill mode in Fig. 9. The "vertical" protein-protein interfaces in Fig. 9 are partly mediated by O-phosphotyrosines (PTR), 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

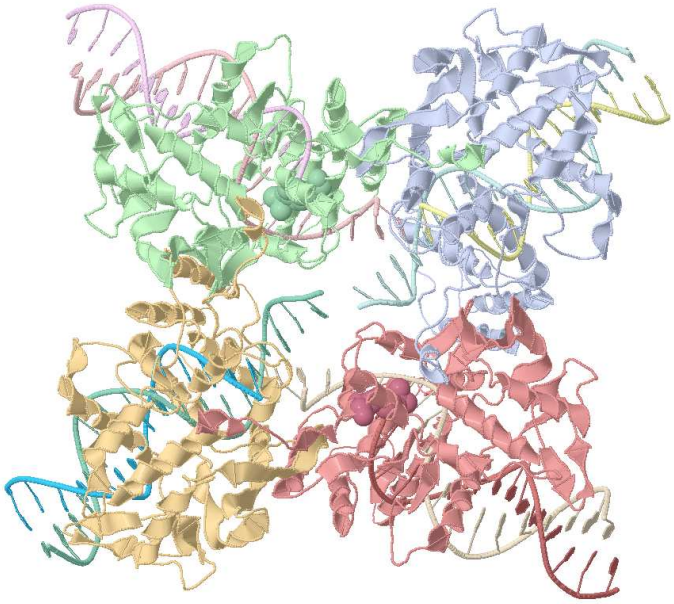

Figure 9: Hetero-dodecamer predicted for PDB entry 1CRX. The complex is made of 4 hetero-trimers, each including a DNA fragment bound to bacteriophage recombinase Cre, considered as primary units in Ref. [94]. The space-fill mode shows O-phosphotyrosine residues (PTR) found in 2 out of 4 protein chains. See discussion in the text. The image was obtained using the Jmol software from `http://www.jmol.org`.

in Fig. 8. Indeed, the structure represents a synapsed complex performing site-specific DNA recombination [94]. The 3-stage reaction opens two DNA strands, initially running across the upper and lower hexamers in Fig. 9, 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 strongest misclassifications in Fig. 3, PDB entry 1TON, is predicted to be a dimeric complex with dissociation barrier of $\Delta G^0_{diss} \approx 37$ 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 above mentioned estimate of about 16 kcal/mol for Zn-protein binding [89]. The value of $\Delta G^0_{diss} \approx 3.2$ 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 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 dataset), for which PISA predicts oligomerization into pipewise structures in presence of Ca and Hg ions, and monomeric states in their absence, in full agreement with experimental findings [96].

The above analysis shows that all 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 definition of 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 Fig. 3, one could use the figure of ± 5 kcal/mol as 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] suggests that the figure may be close to reality. Comparison of entropy change at dissociation of Fasciculin-2 – Acetylcholinesterase complex, 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 ΔG^0_{diss} at room temperatures. 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 stabilization effect of vibrational entropy in flexible complexes.

As follows from the 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 10 demonstrates the scale of uncertainty in the biological unit assignments by comparison of the results obtained from MSD database [98] and

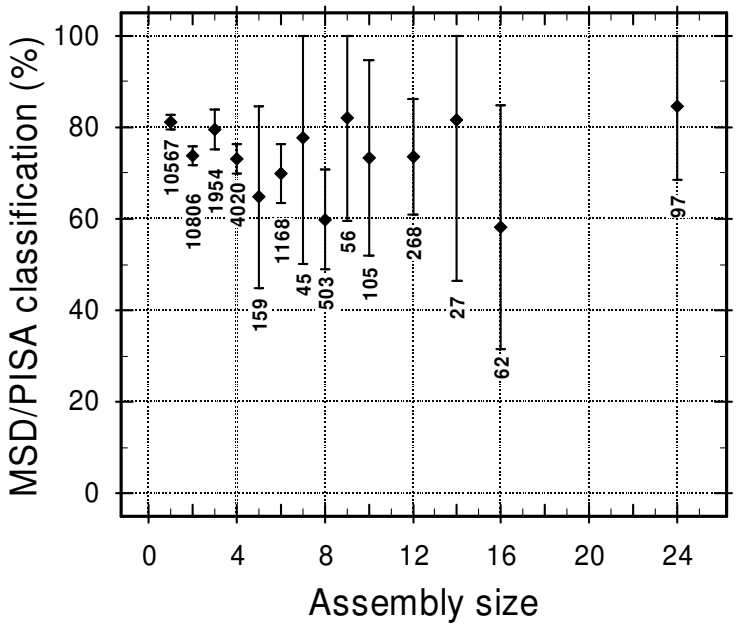

Figure 10: Comparison of multimeric state assignments obtained from MSD database [98] and PISA server at EBI-MSD `http://www.ebi.ac.uk/msd/`. Diamonds denote percentage of coinciding assignments in the respective multimeric class, error bars indicate the 3σ confidence region. Only complete structures 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.

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 Fig. 10, 72% of structures appear to be monomeric or dimeric, where MSD and PISA agree in $81 \pm 1\%$ and $74 \pm 2\%$ 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 dataset above. A simple 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.

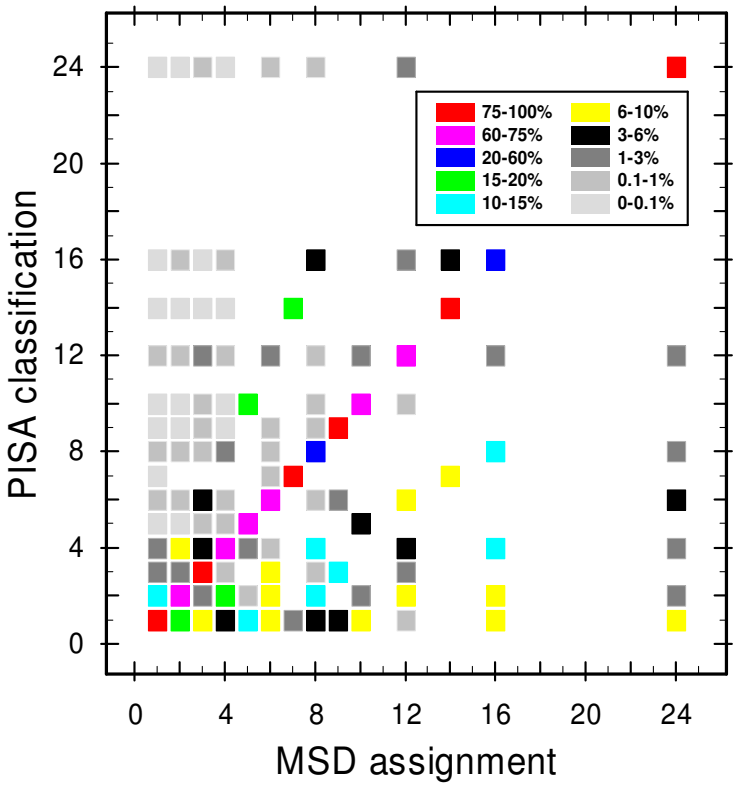

Figure 11: PISA classification of biological unit assignments in the EBI-MSD database [98]. Different colours denote percentage (as shown in the legend) of structures from a particular multimeric class in MSD, classified by PISA into the same or other classes. E.g., 10 to 15% of structures, assigned 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 10.

Figure 11 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 11, most of "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 may appear in PISA as

PDB code	Multimeric state		Ref.
	PISA	Observed	
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]

PDB code	Multimeric state		Ref.
	PISA	Observed	
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 5: Comparison of multimeric states predicted by PISA with those observed experimentally for a set of structures related to bacteriophage T4 studies.

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 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 dataset. We also excluded PDB entry 1MVA (T4 capsid), which is not good for PISA analysis as it does not allow for reconstruction of asymmetric unit because of the absence of NCS records. As may be seen from the Table, PISA analysis failed only in 2 instances (1VQ2 and 1TEO), where hexameric states were

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

Table 6: Comparison of multimeric states assigned by PISA, MSD and PDB to structures representing the strongest PISA misclassifications from Figure 3. 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.

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 5%.

We would like to point out here, that despite enormous curation effort, neither MSD or 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 3, 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.

5 Conclusion

In this paper, we have described 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 (Protein Interfaces, Surfaces and Assemblies), 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, 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 modelled and interpreted, symmetry effects are taken into account, chemical stability of macromolecular complexes is estimated in Gibbs free energy terms along with suggestion of probable dissociation patterns.

On the training dataset, 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 dataset, which includes about 1.35% of all X-ray entries currently in PDB. As found, the strongest misclassifications in the dataset 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 performance of our method beyond the validated training dataset, comparison with partly algorithmic - partly manual annotation in the MSD database [98] shows a reasonable agreement.

In view of the above, we see primary use of PISA as an aid tool for the analysis and modelling of macromolecular interactions, as well as for 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, 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.

Acknowledgement

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. and Fourme, R. The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. J. Mol. Biol. **175** (1984) 159-174.

[2] Berg, J.M., Tymoczko, J.L. and Stryer, L. *Biochemistry.* W.H. Freeman and Co., New York, (2002).

[3] Liu, T. and Chu. B. *Light Scattering by Proteins.* In: Encyclopedia of Surface and Colloid Science, A. Hubbard (ed). Marcel Dekker Inc., New York (2002); pp. 3023-3043.

[4] Feigin L.A. and Svergun D.I. *Structure Analysis by Small Angle X-ray and Neutron Scattering.* New-York: Plenum press, (1987).

[5] Dass, C. *Principles and Practice of Biological Mass Spectrometry.* John Wiley & Sons, Inc., (2001).

[6] Svergun, D.I. and Koch, M.H.J. Advances in structure analysis using small-angle scattering in solution. Cur.Opin.Struct.Biol. **12** (2002) 654-660.

[7] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. The Protein Data Bank. Nucleic Acids Res. **28** (2000) 235–242.

[8] Cavanagh, J., Fairbrother, W.J., Palmer III, A.G. and Skelton N.J. *Protein NMR Spectroscopy.* Academic Press, (1996).

[9] Blundell, T.L. and Johnson, L.N. *Protein Crystallography.* Academic Press Inc. London, (1976).

[10] Ponstingl, H., Henrick, K., and Thornton, J. Discriminating between homodimeric and monomeric proteins in the crystalline state. Proteins **41** (2000) 47–57.

[11] Henrick, K. and Thornton, J. PQS: a protein quaternary structure file server. Trends in Biochem. Sci. **23** (1998) 358–361.

[12] Ponstingl, H., Kabir, T. and Thornton, J. Automatic inference of protein quaternary structure from crystals. J. Appl. Cryst. **36** (2003) 1116–1122.

[13] Jones, S. and Thornton, J.M. Principles of protein-protein interactions. Proc. Natl. Acad. Sci. USA **93** (1996) 13–20.

[14] Argos, P. An investigation of protein subunit and domain interfaces. Protein Eng. **2** (1988) 101–113.

[15] Janin J. and Chothia, C. The structure of protein-protein recognition sites. J. Biol. Chem. **265** (1990) 16027-16030.

[16] Jones, S. and Thornton, J.M. Protein-protein interactions: a review of protein dimer structures. Prog. Biophys. Molec. Biol. **63** (1995) 31–65.

[17] Miller, S. The structure of interfaces between subunits of dimeric and tetrameric proteins. Protein Eng. **3** (1989) 77–83.

[18] Padlan, E.A. 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** (1990) 112–124.

[19] Glaser, F., Morris, R.J., Najmanovich, R.J., Laskowski, R.A. and Thornton, J.M. A method for localizing ligand binding pockets in protein structures. 2006 **62** (Proteins) 479–488.

[20] Gutteridge, A. and Thornton, J.M. Understanding nature's catalytic toolkit. Trends Biochem Sci. **30** (2005) 622–629.

[21] Gutteridge, A., Bartlett, G.J. and Thornton, J.M. Using a neural network and spatial clustering to predict the location of active sites in enzymes. J.Mol.Biol. **330** (2003) 719–734.

[22] Tsai, C.J., Lin, S.L., Wolfson, H. and Nussinov, R. Protein-protein interfaces: Architectures and interactions in protein-protein interfaces and in protein cores. Their similarities and differences. Critical Reviews in Biochem. Mol. Biol. **31** (1996) 127–152.

[23] Lo Conte, L., Chotia, C. and Janin, J. The atomic structure of protein-protein recognition sites. J.Mol.Biol. **285** (1999) 2177–2198.

[24] Chakrabarti, P. and Janin, J. Dissecting protein-protein recognition sites. Proteins **47** (2002) 334–343.

[25] Keskin, O., Tsai, C.J., Wolfson, H. and Nussinov, R. A new, structurally nonredundant, diverse data set of protein interfaces and its implications. Protein Sci. **13** (2004) 1043–1055.

[26] Ogmen, U., Keskin, O., Aytuna A.S., Nussinov, R. and Gursoy, A. PRISM: protein interactions by structural matching. Nucl. Acids Res. **33** (2005) W331–W336.

[27] Preißner, R., Goede, A. and Frömmel, C. Dictionary of interfaces in proteins (DIP). Data bank of complementary molecular surface patches. J.Mol.Bio. **280** (1998) 535–550.

[28] Gong, S., Park, C., Choi, H., Ko, J., Jang, I., Lee, J., Bolser, D.M., Oh, D., Kim, D.S. and Bhak, J. A protein domain interaction interface database: InterPare. BMC Bioinformatics **6** (2005) 207-215.

[29] Moore, W. J. *Physical Chemistry.* Prentice-Hall, Inc., (1972).

[30] Baker, E.N. and Hubbard, R.E. Hydrogen bonding in globular proteins. Prog. Biophys. Molec. Biol **44** (1984) 97–179.

[31] Janin, J., Miller, S., and Chothia, C. Surface, subunit interfaces and interior of oligomeric proteins. J. Mol. Biol. **204** (1988) 155–164.

[32] Horton, N. and Lewis, M. Calculation of the free energy of association for protein complexes. Protein Sci. **1** (1992) 169–181.

[33] Janin, J. and Rodier, F. Protein-protein interaction at crystal contacts. Proteins: Struct. Func. Genet. **23** (1995) 580–587.

[34] Hermann, R. B. Theory of hydrophobic bonding. II. The correlation of hydrocarbon solubility in water with solvent cavity surface area. J. Phys. Chem. **76** (1972) 2754–2759.

[35] Amidon, G. L., Yalkowsky, S. H., Anik, S. T. and Valvani S. C. 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** (1975) 2239–2246.

[36] Floris, F. and Tomasi, J. Evaluation of the dispersion contribution to the solvation energy. A simple computational model in the continuum approximation. J. Comp. Chem. **10** (1989) 616–627.

[37] Born M. Volumes and heats of hydration of ions. Z. Phys. **1** (1920) 45–48.

[38] You T. J. and Bashford D. Conformation and hydrogen ion titration of proteins: a continuum electrostatic model with conformational flexibility. Biophys. J. **69** (1995) 1721–1733.

[39] McDonald I. and Thornton J. Satisfying hydrogen bonding potential in proteins. J. Mol. Biol. **238** (1994) 777–793.

[40] Bahadur R. P., Chakrabarti P., Rodier F. and Janin, J. Dissecting subunit interfaces in homodimeric proteins. Proteins: Struct. Func. Genet. **53** (2003) 708–719.

[41] Xu D., Tsai C.-J. and Nussinov R. Hydrogen bonds and salt bridges across protein-protein interfaces. Protein Engng. **10** (1997) 999–1012.

[42] Pace C., Shirley B., McNutt M. and Gajiwala K. Forces contributing to the conformational stability of proteins. FASEB J. **10** (1996) 75–83.

[43] Fersht A. The hydrogen bond in molecular recognition. Trends Biochem. Sci. **12** (1987) 3214–3219.

[44] Horovitz A., Serrano L., Avron B., Bycroft M. and Fersht A. Strength and co-operativity of contributions of surface salt bridges to protein stability. J. Mol. Biol. **216** (1990) 1031–1044.

[45] Akke M. and Forsen S. Protein stability and electrostatic interactions between solvent exposed charged side chains. Proteins: Struct. Funct. Genet. **8** (1990) 23–29.

[46] Braxton S. *Protein Engineering: Principles and Practice, Cleland J. & Craik C. Eds.*. Wiley-Liss, New York, NY, (1996). Chapter 11

[47] Betz S. Disulphide bonds and the stability of globular proteins. Protein Sci. **2** (1993) 1551–1558.

[48] Clarke J. and Fersht A. Engineered disulfide bonds as probes of the folding pathway of barnase: increasing the stability of proteins against the rate of denaturation. Biochem. **32** (1993) 4322–4329.

[49] Zauhar R. J. and Morgan R. S. A new method for computing the macromolecular electric potential. J. Mol. Biol. **186** (1985) 815–820.

[50] Gilson M. K., Sharp K. A. and Honig B. H. Calculating the electrostatic potential of molecules in solution: method and error assessment. J. Comp. Chem. **9** (1987) 327–335.

[51] Davis M. E. and McCammon J. A. Solving the finite difference linearized Poisson-Boltzmann equation: a comparison of relaxation and conjugaten gradient methods. J. Comp. Chem. **10** (1989) 386–391.

[52] Davis M. E. and McCammon J. A. Electrostatics in biomolecular structure and dynamics. Chem. Rev. **90** (1990) 509–521.

[53] Sharp K. A. and Honig B. Electrostatic interactions in macromolecules: theory and applications. Annu. Rev. Biophys. Biophys. Chem. **19** (1990) 301–332.

[54] Still W. C., Tempczyk A., Hawley R. and Hendrickson R. Semianalytical treatment of solvation for molecular mechanics and dynamics. J. Am. Chem. Soc. **112** (1990) 6127–6129.

[55] Nicholls A. and Honig B. A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson-Boltzmann equation. J. Comp. Chem. **12** (1991) 435–445.

[56] Davis M. E. and McCammon J. A. Dielectric boundary smoothing in finite difference solutions of the Poisson equation: an approach to improve accuracy and convergence. J. Comp. Chem. **12** (1991) 909–912.

[57] Zhou H.-X. Boundary element solution of macromolecular electrostatics: interaction energy between two proteins. Biophys. J. **65** (1993) 955–963.

[58] You T. J. and Harvey S. C. Finite element approach to the electrostatics of macromolecules with arbitrary geometries. J. Comp. Chem. **14** (1993) 484–501.

[59] Purisima E. O. and Nilar S. H. A simple yet accurate boundary element method for continuum dielectric calculations. J. Comp. Chem. **16** (1995) 681–689.

[60] Zaloj V. and Agmon N. Diffusion approach to the linear Poisson-Boltzmann equation. Chem. Phys. Lett. **284** (1998) 76–86.

[61] Bashford D. and Case D. A. Generalized Born models of macromolecular solvation effects. Annu. Rev. Phys. Chem. **51** (2000) 129–152.

[62] Chothia C. Hydrophobic bonding and accessible surface areas in proteins. Nature **248** (1974) 338–339.

[63] Eisenberg, D. and McLachlan, A.D. Solvation energy in protein folding and binding. Nature **319** (1986) 199–203.

[64] Ooi T., Oobatake M., Nemethy G. and Scheraga H. Accessible surface areas as a measure of thermodynamic parameters of hydration of peptides. Proc. Natl. Acad. Sci. USA **84** (1987) 3086–3090.

[65] Vila J., Vasquez M. and Scheraga H. Empirical solvation models can be used to differentiate native from near-native conformations of bovine pancreatic trypsin inhibitors. Proteins: Struct. Funct. Genet. **10** (1991) 199–218.

[66] Wesson L. and Eisenberg D. Atomic solvation parameters applied to molecular dynamics of proteins in solution. Protein Sci. **1** (1992) 227–235.

[67] Juffer A. H., Eisenhaber F., Hubbard S. J., Walther D. and Argos P. Comparison of atomic solvation parametric sets: Applicability and limitations in protein folding and binding. Protein Sci. **4** (1995) 2499–2509.

[68] Wang J. M., Wang W., Huo S. H., Lee M. and Kollman P. A. Solvation Model Based on Weighted Solvent Accessible Surface Area. J. Phys. Chem. B **105** (2001) 5055–5067.

[69] Hou T., Qiao X., Zhang W. and Xu X. Empirical aqueoues solvation models based on accessible surface areas with implicit electrostatics. J. Phys. Chem. B **106** (2002) 11295–11304.

[70] Jackson R. M., Gabb H. A. and Sternberg M. J. E. Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. J. Mol. Biol. **276** (1998) 265–285.

[71] McQuarrie, D.A. *Statistical Mechanics*. New York: Harper & Row, (1976).

[72] Page, M.I. and Jencks, W.P. Entropic contributions to rate accelerations in enzymic and intramolecular reactions and the chelate effect. Proc. Natl. Acad. Sci. USA **68** (1971) 1678–1683.

[73] Murray C. W. and Verdonk, M. L. The consequences of translational and rotational entropy lost by small molecules on binding to proteins. J. Comput.-Aided Mol. Design **16** (2002) 741–753.

[74] Mammen J., Shakhnovich E. I., Deutch J. M. and Whitesides G. M. Estimating the entropic cost of self-assembly of multiparticle hydrogen-bonded aggregates based on the cyanuric acid melamine lattice. J. Org. Chem. **63** (1998) 3821–3830.

[75] Finkelstein A.V. and Janin J. The price of lost freedom: entropy of bimolecular complex formation. Protein Eng. **3** (1989) 1–3.

[76] Kittel C. *Introduction to Solid State Physics*. 7th ed. New York: Wiley, (1995).

[77] Minh D.D.L., Bui J.M., Chang C., Jain T., Swanson J.M.J. and McCammon J.A. The entropic cost of protein-protein association: a case study on Acetylcholinesterase binding to Fasciculin-2. Biophys. J. **89** (2005) L25–L27.

[78] Jaynes E. T. *The Gibbs Paradox.* In: Maximum-Entropy and Bayesian Methods, G. Erickson, P. Neudorfer, and C. R. Smith (eds.). Kluwer, Dordrecht, Holland (1992); pp. 1-22.

[79] Krissinel, E. and Henrick, K. Common subgraph isomorphism detection by backtracking search. Softw. Pract. Exper. **34** (2004) 591–607.

[80] Murakami, M.T., Arruda, E.Z., Melo, P.A., Martinez, A.B., Calil-Eliàs, S., Tomaz, M.A., Lomonte, B., Gutiérrez, J.M., and Arni, R.K. Inhibition of Myotoxic Activity of Bothrops asper Myotoxin II by the Anti-trypanosomal Drug Suramin. J.Mol.Biol. **350** (2006) 416–426.

[81] Antoshenko, T., Min, J.R., Allali-Hassani, A., Dong, A., Weigelt, J., Sundstrom, M., Arrowsmith, C.H., Edwards, A.M., Bochkarev, A. and Plotnikov, A.N. Crystal structure of human sirtuin homolog 5 in complex with Suramin. to be published (2006) .

[82] Evdokimov A.G., Anderson D.E., Routzahn K.M. and Waugh D.S. Unusual molecular architecture of the Yersinia pestis Cytotoxin YopM: a leucine-rich repeat protein with the shortest repeating unit. J. Mol. Biol. **312** (2001) 807–821.

[83] Luscombe N.M., Austin S.E., Berman H.M., and Thornton J.M. An overview of the structures of protein-DNA complexes. Genome Biol. **1** (2000) 1–37.

[84] Krissinel E. and Henrick K. Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Cryst. D **60** (2004) 2256–2268.

[85] Giudice E., Vrnai P. and Lavery R. Base pair opening within B-DNA: free energy pathways for GC and AT pairs from umbrella sampling simulations. Nucl. Acids Res. **31(5)** (2003) 1434–1443.

[86] Daune M. *Molecular biophysics: structure in motion. Part I.* Oxford University Press, New York, (1999).

[87] Rossmann M.G., Mesyanzhinov V.V., Arisaka F. and Leiman P.G. The bacteriophage T4 DNA injection machine. Curr. Opinion Struct. Biol. **14** (2004) 171–180.

[88] Rowsell S., Pauptit R.A., Tucker A.D., Melton R.G., Blow D.M. and Brick P. Crystal structure of carboxypeptidase G_2, a bacterial enzyme with applications in cancer therapy. Structure **5** (1997) 337–347.

[89] DiTusa C.A., Christensen T., McCall K.A., Fierke C.A. and Toone E.J. Thermodynamics of metal ion binding. 1. Metal ion binding by wild-type carbonic anhydrase.. Biochemistry **40** (2001) 5338–5344.

[90] Lubkowski J. and Wlodawer A. Decamers observed in the crystals of bovine pancreatic trypsin inhibitor. Acta Cryst. D **55** (1999) 335–337.

[91] Wu H., Lustbader J.W., Liu Y., Canfield R.E. and Hendrickson W.A. Structure of human chorionic gonadotropin at 2.6 Å resolution from MAD analysis of the selenomethionyl protein. Structure **2** (1994) 545–558.

[92] Lustbader J.W., Yarmush D.L., Birken S., Puett D. and Canfield R.E. The application of chemical studies of human chorionic gonadotropin to visualize its three-dimensional structure. Endocr. Rev. **14** (1993) 291–311.

[93] Littlefield O., Korkhin Y. and Sigler P.B. The structural basis for the oriented assembly of a TBP/TFB/promoter complex. Biochemistry **96** (1999) 13668–13673.

[94] Guo F., Gopaul D.N. and van Duyne G.D. Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. Nature **389** (1997) 40–46.

[95] Fujinaga M. and James M.N.G. Rat submaxillary gland serine protease, tonin structure solution and refinement at 1.8Å resolution. J. Mol. Biol. **195** (1987) 373–396.

[96] Evdokimov A.G., Anderson D.E., Routzahn K.M. and Waugh D.S. Unusual molecular architecture of the Yersinia pestis cytotoxin YopM: a leucine-rich repeat protein with the shortest repeating unit. J. Mol. Biol. **312** (2001) 807–21.

[97] Haramis A.-P. G. and Perrakis A. Selectivity and Promiscuity in Eph Receptors. Structure **14** (2006) 169–171.

[98] Golovin A., Oldfield T.J., Tate J.G., Velankar S., Barton G.J., Boutselakis H., Dimitropoulos D., Fillon J., Hussain A., Ionides J.M.C., John M., Keller P.A., Krissinel E., McNeil P., Naim A., Newman R., Pajon A., Pineda J., Rachedi A., Copeland J., Sitnov A., Sobhany S., Suarez-Uruena A., Swaminathan J., Tagari M., Tromm S., Vranken W. and Henrick K. E-MSD: an integrated data resource for bioinformatics. Nucleic Acids Res. **32** (2004) D211–D216.

[99] Hunt J.F., van der Vies S.M., Henry L. and Deisenhofer J. Structural adaptations in the specialized bacteriophage T4 co-chaperonin Gp31 expand the size of the Anfinsen cage. Cell **90** (1997) 361–371.

[100] Almog R., Maley F., Maley G.F., MacColl R. and van Roey P. Three-Dimensional Structure of the R115E Mutant of T4-Bacteriophage 2'-Deoxycytidylate Deaminase. Biochem. **43** (2004) 13715–13723.

[101] Kanamaru S., Leiman P.G., Kostyuchenko V.A., Chipman P.R., Mesyanzhinov V.V., Arisaka F. and Rossmann M.G. Structure of the cell-puncturing device of bacteriophage T4. Nature **415** (2002) 553–557.

[102] Van Raaij M.J., Schoehn G., Burda M.R. and Miller S. Crystal structure of a heat and protease-stable part of the bacteriophage T4 short tail fibre. J. Mol. Biol. **314** (2001) 1137–1146.

[103] Lariviere L. and Morera S. A base-flipping mechanism for the T4 phage beta-glucosyltransferase and identification of a transition-state analog. J. Mol. Biol. **324** (2002) 483–490.

[104] Leiman P.G., Kostyuchenko V.A., Shneider M.M., Kurochkina L.P., Mesyanzhinov V.V. and Rossmann M.G. Structure of bacteriophage T4 gene product 11, the interface between the baseplate and short tail fibers.. J. Mol. Biol. **301** (2000) 975–985.

[105] Thomassen E., Gielen G., Schutz M., Schoehn G., Abrahams J.P., Miller S. and van Raaij M.J. 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** (2003) 361–373.

[106] Kostyuchenko V.A., Navruzbekov G.A., Kurochkina L.P., Strelkov S.V., Mesyanzhinov V.V. and Rossmann M.G. The structure of bacteriophage T4 gene product 9: the trigger for tail contraction. Struct. Fold Des. **7** (1999) 1213–1222.

[107] Moarefi I., Jeruzalmi D., Turner J., O'Donnell M. and Kuriyan J. Crystal structure of the DNA polymerase processivity factor of T4 bacteriophage. J. Mol. Biol. **296** (2000) 1215-1223.

[108] Papanikolopoulou K., Teixeira S., Belrhali H. Forsyth V.T., Mitraki A. and van Raaij M.J. 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** (2004) 219–227.

[109] Leiman P.G., Shneider M.M., Mesyanzhinov V.V. and Rossmann M.G. Evolution of Bacteriophage Tails: Structure of T4 Gene Product 10. J. Mol. Biol. **358** (2006) 912–921.

[110] Leiman P.G., Shneider M.M., Kostyuchenko V.A., Chipman P.R., Mesyanzhinov V.V. and Rossmann M.G. Structure and location of gene product 8 in the bacteriophage T4 baseplate. J. Mol. Biol. **328** (2003) 821–833.

[111] Raaijmakers H., Toro I., Birkenbihl R., Kemper B. and Suck D. Conformational flexibility in T4 endonuclease VII revealed by crystallography: implications for substrate binding and cleavage. J. Mol. Biol. **308** (2001) 311–323.

[112] Raaijmakers H., Vix O., Toro I., Golz S., Kemper B. and Suck D. X-ray structure of T4 endonuclease VII: a DNA junction resolvase with a novel fold and unusual domain-swapped dimer architecture. EMBO J. **18** (1999) 1447–1458.

[113] Finnin M.S., Cicero M.P., Davies C., Porter S.J., White S.W. and Kreuzer K.N. The activation domain of the MotA transcription factor from bacteriophage T4. EMBO J. **16** (1997) 1992–2003.

[114] Mueser T.C., Jones C.E., Nossal N.G. and Hyde C.C. 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** (2000) 597–612.

[115] Morera S., Imberty A., Aschke-Sonnenborn U., Ruger W. and Freemont P.S. T4 phage beta-glucosyltransferase: substrate binding and proposed catalytic mechanism. J. Mol. Biol. **292** (1999) 717–730.

[116] He M.M., Wood Z.A., Baase W.A., Xiao H. and Matthews B.W. 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** (2004) 2716–2724.

[117] Sickmier E.A., Kreuzer K.N. and White S.W. The crystal structure of the UvsW helicase from bacteriophage T4. Structure **12** (2004) 583–592.

[118] Lariviere L., Gueguen-Chaignon V. and Morera S. 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** (2003) 1077–1086.

[119] Morera S., Lariviere L., Kurzeck J., Aschke-Sonnenborn U., Freemont P.S., Janin J. and Ruger W. High resolution crystal structures of T4 phage beta-glucosyltransferase: induced fit and effect of substrate and metal binding. J. Mol. Biol. **311** (2001) 569–577.

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

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