Detection of protein assemblies in crystals

Evgeny Krissinel and Kim Henrick

European Bioinformatics Institute, Hinxton, Cambridge CB10 1SD UK,
keb@ebi.ac.uk,
WWW: http://www.ebi.ac.uk/msd-srv/prot_int/cgi-bin/piserver

Abstract. The paper describes a new approach to the prediction of probable biological units from protein structures obtained by means of protein crystallography. The method first employs graph-theoretical technique in order to find all possible assemblies in crystal. In second step, found assemblies are analysed for chemical stability and only stable oligomers are left as a potential solution. We also discuss theoretical models for the assessment of protein affinity and entropy loss on complex formation, used in stability analysis.

1 Introduction

Considerable part of protein functionality in biological systems is associated with ability of proteins to bind each other and form stable complexes (assemblies, or biological units). Data on multimeric state of protein complexes and spatial arrangement of their subunits may often provide a deeper insight into the functioning of machinery of life and role of particular proteins in it.

Experimental means for the identification of spatial structure of protein complexes are limited. Because of their relatively large size, protein assemblies are not a good object for NMR studies. Some proteins may exist in dynamic equilibrium between different multimeric states, which also complicates NMR analysis. Electron microscopy is suitable for studying large complexes, but it yields rather low-resolution structures. About 80% of entries in Protein Data Bank [1] represent structures solved by means of X-ray diffraction on protein crystals. In these experiments, crystal structure is identified in the form of atomic coordinates in the asymmetric unit (ASU), unit cell geometry and space symmetry group. However, protein crystallography does not identify true protein associations among all protein contacts in a crystal. At the same time, it is reasonable to expect that protein assemblies do not dissociate during the crystallisation process and therefore protein crystals should contain assemblies as subunits.

Identification of protein assemblies in crystals is, in general, a non-trivial task. The asymmetric unit may be chosen in many different ways, and it does not necessarily coincide with the biological unit. An asymmetric unit may be made from more than one assembly, or a few ASUs may be required to make an assembly, or assembly may be made from several incomplete ASUs. A further complication arises if one assumes that a few different complexes may co-exist in dynamic equilibrium, then crystal may be made from more than one assembly.

Two approaches to the problem have been proposed so far [2, 3]. Both of them are based on the scoring of individual protein interfaces (identified as crystal contacts between monomeric chains) in order to conclude about their biological relevance. PQS server at EBI-MSD [2] scores interfaces mostly on the basis of interface area, with a point system for the hydrophobic effect of complexation, hydrogen bonds, salt bridges and disulphide bonds. The assemblies are built up by the progressive addition of monomeric chains that are bonded by high-scored interfaces. PITA software [3] uses a sophisticated statistical potential to score the interfaces [4] and looks for the solution by iterative bipartitioning of the largest possible assembly in crystal until the minimum-cut interface score exceeds a predefined threshold.

In this paper, we propose another approach based on the consistent enumeration of all assemblies that are possible in a given crystal, with subsequent analysis for chemical stability. The analysis is based on the evaluation of free energy of complex dissociation, which includes the free energy of binding and the entropy change term. As found, the new approach predicts protein assemblies with a higher success rate than its predecessors.

2 Graph-theoretical detection of protein assemblies in crystals

We now want to find all different assemblies in crystal that are allowed by symmetry considerations and content of ASU. We do not assume that crystal is necessarily made from identical assemblies, so that we are looking to find all possible *sets* of different assemblies that fill all the crystal space in a systematic manner. One can note that each such set is unambiguously identified by the inner-assembly interfaces. We will refer to such interfaces as engaged. Then the search may be formulated as enumeration of all possible interface engagements that obey the following rules:

1. Due to crystal symmetry, if an interface of a particular type (that is, between given monomeric chains in a particular relative position) is engaged, all other interfaces of the same type in crystal are also engaged.
2. An interface cannot be engaged if doing so results in assembly that contains identical chains in parallel orientations.

Rule 2 originates from the consideration that if an assembly contains two molecules in parallel orientation, then due to translation symmetry in crystal this assembly must have infinite size. As a consequence of this rule, assembly size cannot exceed the size of unit cell.

The described task may be efficiently addressed by a backtracking scheme, a procedure commonly used in graph matching algorithms [5]. Imagine crystal as a graph where monomeric chains represent vertices, and interfaces between the chains represent edges. The vertices may be calculated by applying all symmetry operations of the required space symmetry group to the chains in ASU, and translating the obtained unit cell according to the cell dimensions and geometry.

```
 1. Calculate periodic graph representing the crystal
 2. List all unique interfaces as I_k
 3. Make empty sets of engaged and tested interfaces {I} := ∅, {T} := ∅
 4. call Backtrack({I}, {T})
 5. stop
procedure Backtrack ( interface sets {I}, {T} )
B.1  copy {T} to {T_1}
B.2  for all interfaces I_k not found in {I} and {T_1} do
B.3    copy {I} to {I_1}
B.4    add I_k to {I_1} and {T_1} (engage interface I_k)
B.5    do
B.6      Identify assemblies formed by interfaces in {I_1}
B.7      Identify induced interfaces and add them to {I_1} and {T_1}
B.8    until no interfaces are induced
B.9    if no assembly contains identical parallel chains then
B.10     output set of assemblies as possible solution
B.11     if more stable assemblies may be found then
B.12       call Backtrack({I_1}, {T_1})
B.13   endif
B.14 done
```

Fig. 1. The assembly enumeration algorithm, see text for details.

The obtained graph is periodic in three dimensions, with period equal to the size of unit cell in the respective dimension. The periodicity allows one to imitate calculations for an infinite crystal on a single unit cell by applying a periodic shift to the inter-cell edges.

The assembly enumeration algorithm is schematically depicted in Fig. 1. It represents a recursive backtracking scheme, which explores all unique combinations of engaged interfaces $\{I\}$. Each such combination corresponds to a set of assemblies which is noted for further analysis of chemical stability. Set $\{T\}$ and its local copies $\{T_1\}$ are used in order to avoid redundant combinations of the interfaces. In steps B.5-8 the algorithm looks for "induced" interfaces and engages them. "Induced" interface is identified as one that appears to be internal to assembly formed by previously engaged interfaces. For example, engaging interfaces $A_1 : A_2$ and $A_2 : A_3$ in trimer (A_1, A_2, A_3) induces interface $A_1 : A_3$.

It may be shown that the total number of unique interface combinations is $N_I!$, where N_I is the total number of unique interfaces. Factorial complexity becomes prohibitive for many PDB entries where $N_I \geq 10$. Therefore, in step B.11 of the algorithm, we terminate those branches of the recursion tree which definitely do not lead to stable assemblies. This technique is borrowed from graph-matching algorithms [5]. The termination condition is derived from the chemical stability analysis and will be described in Section 4.

3 Analysis of chemical stability

Most of assemblies, emerging from the graph-theoretical search, represent unstable structures, which dissociate in dilute solutions. In what follows, we consider assembly as unstable if equilibrtium constant of dissociation is greater than 1. Then protein complex $(A_1, A_2 \dots A_n)$ dissociates into subunits A_i (any subunit may be a multimer) if the free energy change upon dissociation ΔG_{diss} is negative:

$$\Delta G_{diss} = -\Delta G_{int} - T\Delta S < 0 \tag{1}$$

where ΔG_{int} represents free energy of binding of subunits A_i and ΔS is the rigid-body entropy change upon dissociation. Consider terms of Eq. (1) in more detail.

3.1 Free energy of protein binding

The binding energy ΔG_{int} is calculated as a free energy of interface formation between subunits A_i. There are many factors that contribute into protein association energy [6–14], but it is widely acknowledged that major contributions are due to the interaction of protein surface with the solvent and formation of hydrogen bonds and salt bridges across the interfaces:

$$\Delta G_{int} = \Delta G_s(A_1, A_2 \dots A_n) - \sum_{i=1}^{n} \Delta G_s(A_i) - E_{hb}N_{hb} - E_{sb}N_{sb} \tag{2}$$

In Eq. (2), $\Delta G_s(A)$ stands for the solvation free energy of folding. It may be approximated as [11, 16]

$$\Delta G_s(A) = \sum_k \Delta\sigma_k(a_k - a_k^r) \tag{3}$$

where summation is done for all atoms in structure A, a_k stands for the atom's solvent-accessible surface area, $\Delta\sigma_k$ and a_k^r are atomic solvation parameters and surface area in reference state, respectively. $\Delta\sigma_k$ and a_k^r depend on the atom type and charge state in residue. Eq. (2) takes into account that atom charge state may change with changing a_k due to interface formation.

Eq. (2) measures the effect of each of N_{hb} hydrogen bonds and N_{sb} salt bridges between all the subunits A_i by average free energy contributions E_{hb} and E_{sb}, respectively. The strength of a hydrogen bond is estimated to be between 2 and 10 kcal/mol [17]. However, upon disengaging an interface, all potential hydrogen bonding partners become satisfied by hydrogen bonds to water. The only effect that remains here is the decreasing entropy of solvent due to the loss of mobility by bound molecules. Estimations show a contribution of about $E_{hb} \approx 0.6 - 1.5$ kcal/mol per bond [18, 19]. Experimental data on the stabilisation effect of salt bridges are limited. Known studies suggest that free energy contribution of a salt bridge is very close to that of a hydrogen bond $E_{sb} \approx 0.9 - 1.25$ kcal/mol [20, 21].

3.2 Entropy of protein complex formation

Entropy contribution into the free energy of complex dissociation ΔG_{diss} (cf. Eq. (1)) originates from the change of the vibrational mode pattern and regain of rotational and translational degrees of freedom by subunits A_i upon dissociation. Entropy of subunit A may be represented as

$$S(A) = S_{rb}(A) + S_{vib}(A) + S_{surf}(A) \tag{4}$$

where $S_{rb} = S_{trans} + S_{rot}$ stands for the rigid-body (translational and rotational) entropy term, S_{vib} – entropy of internal vibrational modes and S_{surf} – entropy of surface atoms with fractional degrees of freedom.

There are no rigorous theoretical models for the rigid-body entropy of sizeable objects in liquids. Translational entropy contribution S_{trans} may be approximated by the Sackur-Tetrode equation, which was originally derived for the case of small molecules in gas phase [22–24]

$$S_{trans}(A) = R\log\left[\left(\frac{2\pi m(A)kT}{h^2}\right)^{3/2}\left(ve^{5/2}\right)\right] \tag{5}$$

where $m(A)$ is molecular weight and v is the volume open to a molecule. Eq. (5) was found to be a reasonable approximation in liquid phase, too, after corresponding adjustment of the value of v [25].

Rotational rigid-body entropy term can be estimated as [23, 24]

$$S_{rot}(A) = R\log\left[\frac{\sqrt{\pi}}{\sigma(A)}\left(\frac{8\pi^2 kTe}{h^2}\right)^{3/2}\sqrt{J_1(A)J_2(A)J_3(A)}\right] \tag{6}$$

where J_1, J_2 and J_3 are the principle moments of inertia and σ is the symmetry number. This expression seems to be a good approximation in liquids, where rotational entropies were found to differ by only 2% from gas phase values [26].

Vibrational entropy may be estimated as a sum of S_{vib} for all frequences in the molecule's vibration spectra [26]

$$S_{vib} = \sum_k \left[R\frac{hc\nu_k}{kT}\left(\exp\left(\frac{hc\nu_k}{kT}\right) - 1\right)^{-1} - R\log\left(1 - \exp\left(-\frac{hc\nu_k}{kT}\right)\right)\right] \tag{7}$$

where ν_k is kth frequency. Calculation of vibration spectra for protein structures is a computationally hard procedure. As was shown in Ref. [26], usually the value of TS_{vib} is less than 0.5 kcal/mol at normal temperatures, and one can expect that its change at dissociation $T\Delta S_{vib}$ would be much less than that. We therefore neglect vibrational entropy in our model.

The last entropy contribution in Eq. (4), $S_{surf}(A)$, is associated with the mobility of surface (side-chain) atoms. In first approximation, this term may be considered as proportional to the surface area of structure A:

$$S_{surf}(A) = F\sum_k a_k = FW_S(A) \tag{8}$$

where $W_S(A)$ is solvent-accessible surface area of subunit A.

Eqs. (4-8) allow one to estimate a subunit's entropy in solution as

$$S(A) \approx C + \frac{3}{2}R\log\left(m(A)\right) + \frac{1}{2}R\log\left(\frac{J_1(A)J_2(A)J_3(A)}{\sigma^2(A)}\right) + FW_S(A) \qquad (9)$$

This expression contains two empirical parameters: surface entropy factor F, introduced in Eq. (8), and constant entropy term C, which depends on the poorly defined volume v (cf. Eq. (5)). Authors of Ref. [26] estimate uncertainty in S_{trans} as 20-40% of the estimate given by Eq. (5), however state that the expression for S_{rot} (Eq. (6)) is rather precise. We therefore introduce in Eq. (9) the empiric parameter C in attempt to compensate the uncertainty in the definition of v and possibly to account, in first approximation, for other entropy terms, such as conformational entropy, for which no feasible model can be proposed.

Using Eq. (9), entropy change upon complex dissociation in Eq. (1) may be estimated as

$$\begin{aligned}
\Delta S &= \sum_{i=1}^{n} S(A_i) - S(A_1, A_2 \dots A_n) \\
&= (n-1)C + \frac{3}{2}R\log\left(\frac{\prod_i m(A_i)}{\sum_i m(A_i)}\right) + FW_I(A_1, A_2 \dots A_n) \\
&\quad + \frac{1}{2}R\log\left(\frac{\prod_{i;k} J_k(A_i)\sigma^2(A_1, A2 \dots A_n)}{\prod_k J_k(A_1, A_2 \dots A_n)\prod_i \sigma^2(A_i)}\right) \qquad (10)
\end{aligned}$$

where $W_I(A_1, A_2 \dots A_n)$ is buried surface area of subunits A_i in the complex.

3.3 Dissociation pattern

Eqs. (1-3,10) allow one to estimate stability of a protein assembly if its dissociation pattern, or set of subunits $\{A_i\}$, is known. For the purpose of our study it is enough to find at least one dissociation pattern for which $\Delta G_{diss} < 0$ in order to detect instability and to remove the assembly from further consideration.

In order to be a potential dissociation pattern, set of subunits $\{A_i\}$ should satisfy the following conditions:

1. All multi-chain subunits must represent connected stable assemblies.
2. From symmetry considerations, identical interfaces can not be internal to a subuint and separate two subunits in the same dissociation pattern.

Dissociation patterns may be found using a backtracking scheme similar to that shown in Fig. 1. Represent assembly as a graph in which vertices and edges correspond to monomeric chains and interfaces between them, respectively. Then starting point for the algoritm in Fig. 1 would be a non-empty set of all interfaces $\{I\}$ found in assembly (step 3), loop B.2 runs over all interfaces found in $\{I\}$ and not found in $\{T\}$, in steps B.4 and B.7 the algorithm disengages interface I_k and any interfaces induced by that, steps B.9-B.11 are replaced for the calculation

of ΔG_{diss} and stability analysis of the subunits calculated in step B.6. Each subunit is analysed for stability by a recursive application of the bactracking scheme to the subunit. The recursion should terminate once a negative ΔG_{diss} is encountered or all subunits contain only monomeric chains.

Dissociation pattern of stable assemblies may be of a potential interest, too. In general, a protein complex may dissociate in a few different ways, the most efficient of which would be the one with lowest ΔG_{diss}. Dissociation pattern with lowest ΔG_{diss} may be easily identified by the backtracking scheme described above, because it enumerates all possible dissociation patterns for *stable* complexes.

4 Implementation

As described above, our procedure is based on the exchaustive enumeration of all potential assemblies in crystal and their dissociation patterns, using recursive backtracking schemes. Backtracking algorithms are known to be NP-complete and therefore they may be computationally untractable unless a proper termination condition is employed.

Suppose that algorithm in Fig. 1 has generated a set of assemblies that all appear to be unstable, so that $\Delta G^r_{int} + T\Delta S^r > 0$, where index r stands for the recursion level. Then entropy of dissociation on the next level of recursion ΔS^{r+1} should be not less than ΔS^r because any dissociation pattern on level $r+1$ results in the same or larger number of stable subunits than that on level r, while the assembly size only increases with increasing recursion level (cf. Eq. (10)). Maximum energy of binding on level $r+1$ cannot be lower than $\Delta G^r_{int} + \sum_k \Delta G_{int}(I_k)$ where summation is done for all hydrophobic interfaces that still may be engaged, i.e. those with $\Delta G_{int}(I_k) < 0$ and not found in the interface sets $\{I_1\}$ and $\{T_1\}$ (cf. Fig. 1; $\Delta G_{int}(I_k)$ is calculated using Eq. (2) for $n = 2$). Therefore the termination condition is

$$\Delta G^r_{int} + \Delta S^r + \sum_{I_k \notin \{I_1\},\{T_1\}} \min\left(\Delta G_{int}(I_k), 0\right) \geq 0 \qquad (11)$$

where all quantities are calculated for the volume of one unit cell. Despite a very general nature of this estimate, we found that it works very efficiently, especially if interfaces in the backtracking scheme are ordered by increasing $\Delta G_{int}(I_k)$.

In our implementation, we define interface as protein surface area which becomes inaccessible to solvent upon bringing two chains into contact. For the surface area calculations, a method similar to that used in program AREAIMOL of the CCP4 Program Suite [27] was employed. Recipies for the calculation of hydrogen bonds and salt bridges are found in Refs. [6, 15].

Parameters E_{hb}, E_{sb} (cf. Eq. (2)) and C, F (10) were chosen by a fitting procedure using a benchmark set of 218 structures published in Ref. [3]. Since only multimeric states are known for the benchmark structures, we assumed that correct oligomers are the ones of the required multimeric state and lowest ΔG_{diss}

E_{hb}, kcal/mol	E_{sb}, kcal/mol	TC, kcal/mol	TF, kcal/(mol*Å^2)
0.51	0.21	11.7	$0.57 \cdot 10^{-3}$

Table 1. Empirical parameters entering Eqs. (2,10), obtained through the fitting of multimeric states found in the benchmark set of 218 PDB entries from Ref. [3].

(1). Then the parameters were fitted such as to satisfy the following system of inequalities for as many structures as possible:

$$\begin{cases} \Delta G_{diss} > 0 & \text{for correct oligomers} \\ \Delta G_{diss} \leq 0 & \text{for all other multimeric states not lower than the correct one} \end{cases} \tag{12}$$

The described algorithm is implemented as a web-server available at URL given in the title. The server provides pre-calculated data for all PDB entries solved by means of X-ray crystallography, and allows to upload PDB and mmCIF coordinate files for interactive processing. Calculation time depends drastically on the number of different interfaces in crystal, however most of entries are solved in a few-minute time. The server also provides a detail annotation of interfaces and structures, visualisation of assemblies and database search tools.

5 Results and discussion

The resulting values of empirical parameters, used in Eq. (10), are listed in Table 1. As seen from the Table, energy effect of hydrogen bonds and salt bridges appears to be somewhat smaller than the estimates given in the above discussion, but well within a reasonable range. Given that significant interfaces normally have 10-20 and more hydrogen bonds, their contribution to the free energy of binding G_{int} appears to be comparable with that of hydrophobic interactions. Entropy contribution from the frozen motion of surface atoms in interfaces, F, is quite small, just over 0.5 kcal/mol per 10^3Å^2 of interface area. Most of entropy change at complex formation comes from the constant entropy term, C, followed by the mass- and moment of inertia- dependent terms (cf. Eq. (10)). Mathematically, the system of inequalities (12) appears slightly underfit, which means that the used benchmark set may be insufficient for the calibration purposes, and the results may still be improved if a larger data set is used.

Table 2 presents the assembly classification results obtained for the benchmark set of 218 PDB entries [3], used for the calibration of empirical parameters. Each row of the Table corresponds to one of 5 oligomeric classes present in the benchmark set, and columns give the classification counts obtained for that class. As seen from the Table, we have obtained a nearly uniform success rate across different oligomeric classes, with the lowest rate of 87% for tetramers. Tetramers have also been found as the least predictable oligomeric class in Ref. [3], with considerably larger differences between the classes. The overall success rate is 90%, which is higher than the one reported in Ref. [3] (84%). On comparison,

	1mer	2mer	3mer	4mer	6mer	Other	Sum	Correct
1mer	50	4	0	1	0	0	55	91%
2mer	6	68+11	0	2+1	0	0	76+12	90%
3mer	1	0	22	0	1	0	24	92%
4mer	2	3	0	27+6	0	0	32+6	87%
6mer	0	0	0	1	10+2	0	11+2	92%
						Total:	198+20	90%

Table 2. Assembly classification obtained for the benchmark set of 218 PDB entries from Ref. [3]. 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.

the PQS server at EBI-MSD gives 78% of correct answers, however this figure is less indicative because PQS was not optimised for the used benchmark set.

A detail study of misclassified cases shows a typical misestimate of ΔG_{diss} (Eq. (1)) within ± 5 kcal/mol. This value could be taken as a precision limit for the models proposed in Section 3 if multimeric states in the benchmark set are trusted. There is, however, one example of misclassification that is far beyond any reasonable precision range for the method. PDB entry `1qex` contains two identical chains, which should form a homo-trimer [3]. Our procedure, as well as PQS [2], suggests that it is actually a homo-hexamer shown in Fig. 2. Calculation results indicate that the most favourable dissociation pathway for this assembly is through a detachement in the isthmus between the two identical trimers with $\Delta G_{diss} \approx 90$ kcal/mol. Such high value of the dissociation barrier implies that the structure could well be hexameric.

The example of `1qex` may indicate that not all multimeric states given in the used benchmark set are correct. A probable source of errors may be that only one oligomer from a few of them in chemical equilibrium is reliably detected in experiment. However, we tend to explain most of misclassifications by neglecting the specific experimental conditions, such as concentration, pH, tem-

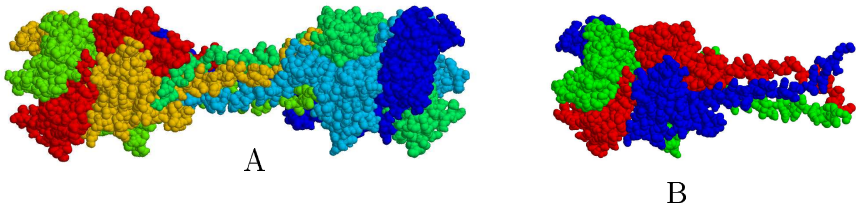

Fig. 2. Homo-hexamer found for PDB entry `1qex` (A), and homo-trimer (B) which should be the correct multimeric state according to data in Ref. [3], see discussion in the text. The images were obtained using the Rasmol software [28].

	1mer	2mer	3mer	4mer	5mer	6mer	8mer	10mer	12mer	Sum	Correct
1mer	131	11	0	4	0	2	2	0	0	150	87%
2mer	12+6	88+12	0+1	4	0	1	0+2	0	0	105+21	79%
3mer	1	0+2	6+2	0	0	0+1	0	0	0	7+5	66%
4mer	1+1	5+2	0	25+5	0	0	1+2	0	0	32+10	71%
5mer	0+1	0	0	0	2+1	0	0	0	0	2+2	75%
6mer	0+1	2+1	0	0	0	13+2	0	0	0	15+4	79%
8mer	0	1	0	0	0	0	0+2	0	0	1+2	67%
10mer	0	0	0	0	0	0	0	2	0	2	100%
12mer	2	0	0	0	0	0	0	0	5+1	7+1	75%
									Total:	321+45	81%

Table 3. Assembly classification obtained for the new entries deposited into PDB through EBI-MSD deposition site. The reference classification has been done in MSD by manual curation. See Table 2 for used notations.

perature and presence of other agents, in our models. A thorough account of all affecting factors is difficult and if done then requires a quite detail description of experimental conditions from a user.

Most structures are deposited into PDB without experimental evidence of their oligomeric states. The benchmark set of 218 PDB entries published in Ref. [3] contains all structures with oligomeric states that are currently known to us as experimentally verified. Biological unit assignments in PDB is based mainly on the curators' scientific experience. Table 3 compares automatic assembly classification, obtained by us, with manual curation results for 366 new entries deposited recently into PDB at the EBI-MSD deposition site. As seen from the Table, most (75%) of the depositions were classified as monomers and dimers, which is reproduced at 87% and 79% success rate, respectively. Success rate for other oligomeric classes varies from 66% to 100%, however these figures are less indicative because of too few structures present. Overall, 81% of automatic an manual classifications agree with each other.

The most frequent misclassifications in Table 3 are dimers instead of tetramers, then monomers instead of dimers and vice versa. These are special cases when a larger assembly may or may not be divided in two parts. A detail study of the misclassifications reveals that in most of them ΔG_{diss} lies within ± 5 kcal/mol, the same uncertainty as that found for the benchmark set. A few strongest exceptions to this observation are shown in Table 4. Visual inspection of these assemblies reveals a poor packing quality of their interfaces (except for well-packed `1y6x` and `1y7p`), which fact could suggest classification into lower oligomeric classes. However, our calculations show that, despite their topological imperfectness, the interfaces represent pronounced hydrophobic patches. This means that the interfaces may be stronger than visually appears, which makes higher oligomeric states possible. A definite answer as to what the oligomeric state actually is in these cases, as well as in cases with low $|\Delta G_{diss}|$, may be given only by experimental study.

PDB entry	1y6x	1ywk	1v7y	1wq5	2bh8	1y7p	1ylf
Assigned state	1mer	1mer	1mer	1mer	4mer	2mer	1mer
Calculated state	4mer	6mer	2mer	2mer	8mer	6mer	6mer
ΔG_{diss}, kcal/mol	16.5	9.9	9.0	15.3	9.2	36.1	16.2

Table 4. The strongest misclassifications in Table 3. See text for details.

6 Conclusion

We have described here a novel method for the calculation of biological units from protein crystallography data. In difference of its predecessors, our method is based on the stability analysis of all assemblies allowed by crystal symmetry and geometry of unit cell. We estimate the free energy of dissociation using theoretical models for free energy of protein binding and rigid-body entropy of protein assemblies. This approach allows us not only to predict the multimeric states and 3D arrangements of monomeric units with 80-85% accuracy, but also to guess on the probable dissociation patterns of assemblies.

The described procedure is implemented as a web-server available at URL given in the title of this paper. The server provides a detail summary of all crystal contacts and monomeric chains, list of probable protein assemblies, as well as searching for alike interfaces in the PDB archive.

Although our models neglect specific conditions, such as concentration and pH, which may affect formation of assemblies, predictive power of the method appears to be sufficiently high. Further studies are needed to improve the theoretical models of protein affinity and entropy change upon assembly formation.

Acknowledgement
E.K. is supported by the research grant No. 721/B19544 from the Biotechnology and Biological Sciences Research Council (BBSRC) UK. The authors thank Mr. A. Hussain for his work on the comparison of assigned and calculated oligomeric states using the described software.

References

1. 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.
2. Henrick, K. and Thornton, J.: PQS: a protein quaternary structure file server. Trends in Biochem.l Sci. **23** (1998) 358–361.
3. Ponstingl, H., Kabir, T. and Thornton, J.: Automatic inference of protein quaternary structure from crystals. J. Appl. Cryst. **36** (2003) 1116–1122.
4. Ponstingl, H., Henrick, K., and Thornton, J.: Discriminating between homodimeric and monomeric proteins in the crystalline state. Proteins **41** (2000) 47–57.
5. Krissinel, E. and Henrick, K.: Common subgraph isomorphism detection by backtracking search. Softw. Pract. Exper. **34** (2004) 591–607.

6. Baker, E.N. and Hubbard, R.E.: Hydrogen bonding in globular proteins. Prog. Biophys. Molec. Biol **44** (1984) 97–179.
7. Janin, J., Miller, S., and Chothia, C.: Surface, subunit interfaces and interior of oligomeric proteins. J. Mol. Biol. **204** (1988) 155–164.
8. Argos, P.: An investigation of protein subunit and domain interfaces. Protein Eng. **2** (1988) 101–113.
9. Miller, S.: The structure of interfaces between subunits ofdimeric and tetrameric proteins. Protein Eng. **3** (1989) 77–83.
10. Janin J. and Chothia, C.: The structure of protein-protein recognition sites. J. Biol. Chem. **265** (1990) 16027-16030.
11. Horton, N. and Lewis, M.: Calculation of the free energy of association for protein complexes. Protein Sci. **1** (1992) 169–181.
12. Janin, J. and Rodier, F.: Protein-protein interaction at crystal contacts. Proteins: Struc. Func. Genet. **23** (1995) 580–587.
13. Jones, S. and Thornton, J.M.: Protein-Protein interactions: a review of protein dimer structures. Prog. Biophys. Molec. Biol. **63** (1995) 31–65.
14. Jones, S. and Thornton, J.M.: Principles of protein-protein interactions. Proc. Natl. Acad. Sci. USA **93** (1996) 13–20.
15. Xu, D., Tsai, C.-J. and Nussinov, R.: Hydrogen bonds and salt bridges across protein-protein interfaces. Protein Engng. **10** (1997) 999–1012.
16. Eisenberg, D. and McLachlan, A.D.: Solvation energy in protein folding and binding. Nature **319** (1986) 199–203.
17. McDonald I. and Thornton J.: Satisfying hydrogen bonding potential in proteins. J. Mol. Biol. **238** (1994) 777–93.
18. Pace C., Shirley B., McNutt M. and Gajiwala K.: Forces contributing to the conformational stability of proteins. FASEB J. **10** (1996) 75–83.
19. Fersht A.: The hydrogen bond in molecular recognition. Trends Biochem. Sci. **12** (1987) 3214–3219.
20. Horovitz A., Serrano L., Avron B., Bycroft M. and Fersht A.: Strength and cooperativity of contributions of surface salt bridges to protein stability. J. Mol. Biol. **216** (1990) 1031–1044.
21. Akke M. and Forsen S.: Protein stability and electrostatic interactions between solvent exposed charged side chains. Proteins: Struct. Funct. Genet. **8** (1990) 23–29.
22. 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.
23. McQuarrie, D.A. *Statistical Mechanics*. New York: Harper & Row, (1976).
24. 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.
25. Finkelstein, A.V. and Janin, J.: The price of lost freedom: entropy of bimolecular complex formation. Protein Eng. **3** (1989) 1–3.
26. 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 AcidMelamine Lattice. J. Org. Chem. **63** (1998) 3821–3830.
27. Collaborative Computational Project, Number 4.: The CCP4 Suite: Programs for Protein Crystallography. Acta Cryst. D **50** (1994) 760-763.
28. Sayle, R. A., and Milner-White, E. J.: RasMol: Biomolecular graphics for all. Trends in Biochemical Sci. **20** (1995) 374-376.