CRK: An evolutionary approach for distinguishing biologically relevant interfaces from crystal contacts

Martin A. Schärer,[1,2] Markus G. Grütter,[2] and Guido Capitani[1,2*]

[1] Biomolecular Research, Paul Scherrer Institut, Villigen CH-5232, Switzerland
[2] Biochemisches Institut der Universität Zürich, Zürich CH-8057, Switzerland

ABSTRACT

Protein crystals contain two different types of interfaces: biologically relevant ones, observed in protein–protein complexes and oligomeric proteins, and nonspecific ones, corresponding to crystal lattice contacts. Because of the increasing complexity of the objects being tackled in structural biology, distinguishing biological contacts from crystal contacts is not always a trivial task and can lead to wrong interpretation of macromolecular structures. We devised an approach (CRK, core-rim K_a/K_s ratio) for distinguishing biologically relevant interfaces from nonspecific ones. Given a protein–protein interface, CRK finds a set of homologs to the sequences of the proteins involved in the interface, retrieves and aligns the corresponding coding sequences, on which it carries out a residue-by-residue K_a/K_s ratio (ω) calculation. It divides interface residues into a "rim" and a "core" set and analyzes the selection pressure on the residues belonging to the two sets. We developed and tested CRK on different datasets and test cases, consisting of biologically relevant contacts, nonspecific ones or of both types. The method proves very effective in distinguishing the two categories of interfaces, with an overall accuracy rate of 84%. As it relies on different principles when compared with existing tools, CRK is optimally suited to be used in combination with them. In addition, CRK has potential applications in the validation of structures of oligomeric proteins and protein complexes.

Proteins 2010; 78:2707–2713.
© 2010 Wiley-Liss, Inc.

Key words: structural bioinformatics; protein–protein interactions; K_a/K_s; selection pressure.

INTRODUCTION

In protein crystallography, contacts between molecules can be divided into two categories: biologically relevant interfaces and crystal contacts. While the former represent interfaces belonging to oligomeric proteins or protein–protein complexes, the latter are necessary for the formation of the crystal lattice. With the increasing complexity of the biological macromolecules and complexes being currently tackled, distinguishing crystal contacts from biological interfaces can be, more often than commonly thought, a far from trivial task. In fact, the physicochemical features, like H-bonding networks and hydrophobic interactions, of the two kinds of interfaces are very similar. Also, crystal contacts are often quite extensive, making a visual assessment of the oligomeric state of a protein from its crystal assembly unreliable. If no other biochemical evidence is accessible, for example, in cases where structural determination precedes biochemical characterization, bioinformatic approaches are required that are able to predict the true oligomeric state of the protein or the biological relevance of a contact observed between two or more putative interaction partners. Such situations are becoming increasingly common in the context of structural genomics projects.

The crystal structure of complex $\mathrm{FimD_N}$-FimC-$\mathrm{FimH_P}$[1] from the type 1 pilus of uropathogenic *E. coli* provides a very good example of this difficulty. The asymmetric unit of the crystals contains a heterotrimeric protein complex that could be interpreted in two radically different ways depending on which of the observed contacts were considered to be the biological ones. Biochemical data indicated that $\mathrm{FimD_N}$ recognizes a binary complex of FimC with $\mathrm{FimH_P}$ but not unbound FimC alone. Based on that finding, the arrangement featuring the larger interface area (albeit small in absolute terms) between $\mathrm{FimD_N}$ and FimH [$\mathrm{FimD_N}$-FimC-FimH (red-cyan-yellow) in Fig. 1] seemed more likely. Further biochemical analysis, however, led to a counterintuitive conclusion: a $\mathrm{FimD_N}$ mutant lacking the first 24 N-terminal residues proved incapable

Additional Supporting Information may be found in the online version of this article.
Abbreviations: ASA, accessible surface area; BSA, buried surface area; CRK, core-rim K_a/K_s ratio; LCA, lower boundary for core assignment; ω, K_a/K_s ratio.

Grant sponsors: Forschungskommission of the Paul Scherrer Institut, NCCR Structural Biology program (Schweizerischer Nationalfonds, ETH Zurich and University of Zurich)
*Correspondence to: Guido Capitani, Biomolecular Research, Paul Scherrer Institut, OFLC/110, 5232 Villigen, Switzerland. E-mail: guido.capitani@psi.ch.
Received 25 February 2010; Revised 6 May 2010; Accepted 18 May 2010
Published online 1 June 2010 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22787

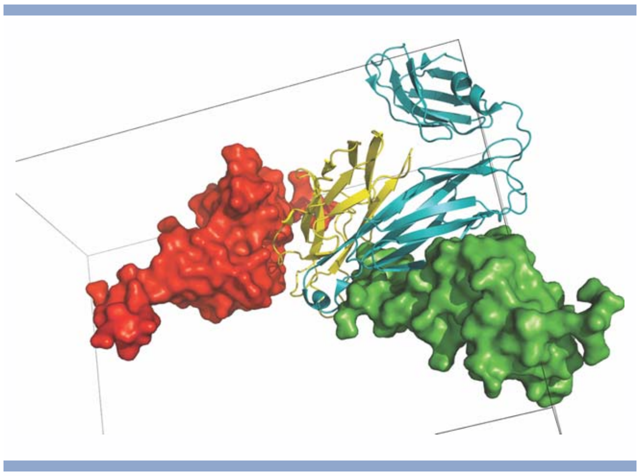

Figure 1

Depiction of two possible assemblies of the ternary complex of FimD_N (green, red), FimC (cyan), and FimH_P (yellow) from the *E. coli* type 1 pilus in the crystal lattice of PDB entry 1ZE3: green-cyan-yellow (biologically relevant interface) or red-cyan-yellow (crystal contact). It is known from biochemical characterization that FimD_N does not recognize FimC alone. Interface areas, $\text{FimD}_{N,\text{red}}$-$\text{FimH}_P$: 486 $Å^2$; $\text{FimD}_{N,\text{green}}$-$\text{FimH}_P$: 209 $Å^2$; $\text{FimD}_{N,\text{green}}$-FimC: 1078 $Å^2$. Figure is generated with PyMOL.[23]

of binding the binary complex of FimC with FimH_P. Those 24 residues contribute only to the assembly with the smaller FimD_N to FimH interface [FimD_N-FimC-FimH (green-cyan-yellow) in Fig. 1], thus identifying it as the biologically relevant one.

Several computational approaches for distinguishing biological contacts from crystal contacts have been published. What follows is a nonexhaustive list of available methods. The first one, introduced by Janin, relied on the size statistics of crystallographic interfaces. The author derived a formula to calculate the probability of finding a nonspecific interface based on the size of the interface.[2]

In 2004, Bahadur *et al.* proposed a method in which interface residues were divided into a "rim" and a "core" set, depending on their degree of burial on interface formation (core residues get completely buried on interface formation, while rim residues get only partially buried). The authors observed that crystal contacts possess a lower amount of core residues than biologically relevant interfaces.[3]

In 2001, Elcock and McCammon devised a method based on the comparison of the sequence entropy (derived by multiple sequence alignments) of interface residues versus that of noninterface residues. The underlying assumption that the average sequence entropy of interface residues ($<s>_{\text{interf}}$) should be lower than that of noninterface residues ($<s>_{\text{noninterf}}$), so their ratio be smaller than 1, was confirmed by numerical tests.[4]

To avoid possible bias by low-entropy noninterface residues, such as those in exposed active sites or allosteric binding sites, Guharoy and Chakrabarti[5] modified the method of Elcock and McCammon by restricting the sequence entropy analysis to interface residues only and applying the rim-core concept developed by Bahadur *et al.* Interface residues were divided into a rim and a core set, and then the average entropies of the two sets were compared under the assumption that for a biologically relevant interface, $<s>_{\text{core}}$ be lower than $<s>_{\text{rim}}$, so their ratio be smaller than 1.

A completely different, sequence-independent approach, called PISA (Protein Interfaces, Surfaces and Assemblies), successfully relies on thermodynamic stability calculations[6,7] and is implemented in a web server. A further method, PITA (Protein InTerfaces and Assemblies), analyzes contacts based on their size and chemical complementarity.[8]

In another attempt to classify protein interaction types, NOXclass[9] uses a support vector machine integrating six different properties (geometric and sequence based) to distinguish three categories of contacts (biological obligate, biological nonobligate, and crystal contacts).

Recently, Bernauer *et al.*[10] have presented DiMoVo, a method based on Voronoi tessellation. It also relies on a support vector machine for parameter optimization. The issue of distinguishing crystal from biological contacts has been reviewed, with practical examples, by Kobe *et al.*[11]

The approach proposed in this article, called core-rim K_a/K_s ratio (CRK), makes use of a modified core-rim subdivision of interface residues and moves the analysis from amino acid level to coding sequence level by analyzing per-residue selection pressure estimates. The rationale for the method is as follows: residues involved in interface formation are subject to a selection pressure, which is assumed to be more stringent in core residues than in rim residues. We use K_a/K_s ratio (ω) values, calculated by SELECTON,[12,13] as metrics for selection pressure and combine the scores with a geometric description of the interface. The approach provides an effective method to distinguish crystal contacts from biologically relevant interfaces.

MATERIALS AND METHODS

The CRK approach was implemented for the Linux operating system using the perl scripting language and bash scripts. To carry out some specific steps, the method uses various established bioinformatic applications: BLAST[14] for similarity searches; T-COFFEE[15] for multiple sequence alignment; and SELECTON[13] for the calculation of ω values.

As the first step of the CRK calculations, the structure corresponding to the input PDB code, if not already present in the working directory, is retrieved from the PDB.[16] The sequence of each protein chain in the file is extracted and used as a query for a BLASTP search for homologs versus the UniProt database (version 15.14).[17]

The resulting list of putative homologues is then filtered according to the following criteria:

a. Identity, with threshold given by the user (default = 60%): In the case in which fewer than 10 suitable homologs are found, the identity threshold can be relaxed stepwise (in steps of 1% identity) till a specified plateau level.
b. Taxonomy: The taxonomic domain (eukaryotes, prokaryotes, archaea, viruses) is parsed from the query Uniprot OC field. Alternatively, the search for homologs can be restricted to any taxonomic group listed in the OC record.
c. A redundancy reduction procedure limits the overall number of sequences to 60.

The coding sequence for each entry, as well as for the query, is obtained from the EMBLCDS database and aligned based on a T-COFFEE protein multiple sequence alignment (MSA). If no EMBLCDS sequence corresponding to the query is available from EMBLCDS, the program aborts. The program also stops if less than 10 coding sequences are obtained.

The resulting multiple sequence alignment of coding sequences is then analyzed with SELECTON for the selection pressure on each site.

The geometric description of the interface, currently retrieved automatically from PISA[7] (or manually, by user's choice, i.e., in case of a nondeposited structure), is used to divide the interface residues into a rim and a core subset according to the percentage of their accessible surface area (ASA) that gets buried on interface formation. The interface description from PISA is used only to obtain the list of residues participating in the interface and their buried surface area (BSA) and can in principle be substituted by the output of any other algorithm that lists BSA values of an interface. Each interface residue is attributed an ω value, and the arithmetic and weighted averages are calculated according to the formulas:

$$\langle\omega\rangle_{\mathrm{core}} = \frac{\sum_i \omega(i)}{N} \quad \text{and} \quad \langle\omega\rangle_{\mathrm{rim}} = \frac{\sum_j \omega(j)}{M}$$

$$\langle\omega\rangle_{\mathrm{corew}} = \frac{\sum_i \omega(i)\mathrm{BSA}(i)}{\sum_i \mathrm{BSA}(i)} \quad \text{and} \quad \langle\omega\rangle_{\mathrm{rimw}} = \frac{\sum_j \omega(j)\mathrm{BSA}(j)}{\sum_j \mathrm{BSA}(j)}$$

where N and M are the number of core and rim residues, and i and j are the core and rim residues, respectively.

Finally, the two ratios CRK and wCRK for the interface under investigation are computed:

$$\mathrm{CRK} = \frac{\langle\omega\rangle_{\mathrm{core}}}{\langle\omega\rangle_{\mathrm{rim}}} \quad \text{and} \quad \mathrm{wCRK} = \frac{\langle\omega\rangle_{\mathrm{corew}}}{\langle\omega\rangle_{\mathrm{rimw}}}$$

A biological interface is expected to possess a CRK value smaller than 1 as the core of the interface is under stronger purifying selection than the rim. The cutoff of significance (S) was set, following extensive testing (see Supporting Information Materials and Methods), at 0.85. An interface for which CRK and/or wCRK is higher than S is considered as a crystal contact. If an interface possesses less than a minimal number of core residues (N, default $N = 6$) under the initial burial percentage threshold for core residue assignment CA (burial percentage on interface formation $= 100 \times \mathrm{BSA/ASA}$, default value = 95%), CA is relaxed in steps of 1% till $N \geq 6$ or CA drops below a lower boundary for core assignment (LCA, default = 82%). In the latter case, the interface is considered as a crystal contact for whatever CRK or wCRK values.

Third party tools used in CRK

BLAST

BLAST version 2.2.18 is installed locally and used with low-complexity filter off. The database used for the search is Uniprot (version 15.14).

T-COFFEE

The multiple sequence alignment is performed using standalone version 2.66.

SELECTON

Selection pressure is estimated using a locally installed version (2.4) of SELECTON, with a precision level $\epsilon = 0.05$, the M8 codon substitution model,[18] and the Bayesian method. The minimum default number of sequences required is 10.

PISA server

The output (interface description) from the server implementation is used (version 1.18).

Datasets

For the first evaluation of CRK and for parameter search, three datasets were used: (1) nine "difficult" dimeric proteins[4]; (2) pyridoxal phosphate (PLP)-dependent enzyme structures derived from a list compiled by Percudani and Peracchi[19] and extended with an additional search against the PDB (on 2009/06/12); and (3) the DiMoVo set of crystal monomers and dimers.[10]

The final values for three CRK parameters, S (cutoff of significance), N (minimal number of core residues), and LCA ($S = 0.85$, $N = 6$, LCA = 82%), were determined as described in Supporting Information Text and in Supporting Information Figures 1 and 2.

The performance of CRK was tested and benchmarked against a high-resolution dataset (HRset) and a dataset of dimeric interfaces derived from Dey *et al.*[20]

Table I
CRK Values for the Set of "Difficult" Dimeric Proteins Compared with the Results of Elcock and McCammon (Indicated as E & M) and the Results of Guharoy and Chakrabarti (Indicated as G & C)

Interface	Area (Å^2)	CA (%)	N	CRK	wCRK	E & M	G & C
1A3C_1	989	95	14	0.24	0.23	0.90	0.60
1AFW_1[a]	2,396	95	34	0.55	0.51	0.99	0.83
1ALK_1	3,824	95	42	0.35	0.45	0.97	0.90
1AOR_1	1,232	95	12	0.36	0.22	1.49	1.07
1CZJ	n.a.	n.a.	n.a.	n.a.	n.a.	1.09	1.01
1ICW_1	990	95	10	0.58	0.71	1.15	0.66
1PRE_1	2,240	85	6	0.66	0.77	1.09	0.79
1SMT_1[a]	1,974	95	15	0.53	0.41	0.98	0.95
2TCT_1	2,612	95	11	0.48	0.60	1.07	1.07

1CZJ: not available (n.a.) because not enough coding sequences related to the query were found; 1PRE: values after manual correction of some sequence differences to weight. For the previous methods, values lower than 1 indicate a biological interface, whereas for our method, CRK and wCRK need to be lower than 0.85.

[a]The identity threshold was allowed to relax in steps of 1% till at least 10 entries were retrieved or an identity plateau of 50% was reached. *N* indicates the number of core residues, and CA indicates the core assignment threshold.

The HRset was obtained by retrieving from the PDB (on January 29, 2010) all crystal structures with a resolution of at least 1.0 Å and experimental data deposited. The dataset was further filtered according to the "macromolecule type" record of the entries (protein only) and to their source organism record (cellular organism). In addition, only entries sharing a sequence identity level not higher than 50% were retained. The HRset was introduced as a means to compare the performance of the CRK indicator calculated on the datasets of heterogenous quality (the other datasets) to a set of high-quality structures, where the side chains can be assumed to represent a "native" conformation.

RESULTS AND DISCUSSION

The CRK indicator is computed based on a search for homologous sequences that are then retrieved and used in a multiple sequence alignment. The corresponding coding sequences are retrieved as well and aligned based on that MSA. The MSA of DNA sequences is then analyzed for selection pressure, and the calculated ω values are assigned to their respective interface residues to calculate the CRK parameter.

A set of nine "difficult" dimeric proteins (Dataset 1)

This small set of proteins was compiled by Elcock and McCammon[4] while benchmarking their method on a larger set. It consists of dimeric proteins (1PRE and 1ICW are nonobligate dimers) with sequence entropy features making them difficult to assign as dimeric. This set was also used by Guharoy and Chakrabarti[5] for the same purpose.

The CRK method correctly identified eight of the nine proteins of this set as dimers (Table I). One entry, 1CZJ (a class III cytochrome *c*), could not be analyzed as no corresponding coding sequence was found in the database. As they are not bound to the number of available coding sequences, both the Elcock and McCammon and the Guharoy and Chakrabarti methods are able to produce a result for entry 1CZJ.

On running CRK on 1PRE, proaerolysin, it became apparent that the deposited PDB file contained some sequence differences when compared with the EMBLCDS entry. It was thus necessary to change a few residues of the protein FASTA sequence to fix the problem. If such a problem arises, the current implementation of CRK provides the user with a point-by-point guide to manually solve it. For 1AFW and 1SMT (the footnote "a" in Table I), not enough entries were available with an identity threshold of 60%, and the program was allowed to relax the threshold in steps of 1% till reaching a plateau of 50% identity.

Thus, the accuracy of CRK was 100% on this dataset, with a recall of 88.9% (Table I).

For the runs described in the following paragraphs, a strict identity threshold of 60% was kept.

PLP-dependent enzymes (Dataset 2) and the DiMoVo set (Dataset 3)

As a second dataset, we used a list of PLP-dependent enzymes compiled by Percudani and Peracchi[19] to find the corresponding structures in the PDB. The quaternary structure of such enzymes is known to be dimeric or higher oligomeric (multiples of dimers),[21] and therefore, they represent a safe benchmark for the detection of biologically significant interfaces. The list was extended with a search for further PLP enzyme structures in the PDB.

For the 61 PDB codes, 104 interfaces were collected from the PISA server. For 19 interfaces, not enough homologous sequences meeting the selection parameters could be found. Of the 81 interfaces analyzed, 74 were correctly assigned as biologically relevant, with an accuracy of 91.4% and a recall of 71.2%. All results for the PLP enzyme dataset are reported in Supporting Information Table I.

As third dataset, we used the DiMoVo set of dimers and crystal monomers.[10] For the dimer subset, encompassing 36 interfaces (see Supporting Information Text), the CRK tool calculated 27 and correctly assigned 22 interfaces, with an accuracy of 81.5% and a dimer recall of 61.1%. For the monomer subset, encompassing 137 interfaces, 78 were calculated and 68 correctly assigned as crystal monomers (accuracy 87.2% and monomer recall of 49.6%). The overall accuracy for the DiMoVo set was 85.7% with a recall of 52.0%. All results for the DiMoVo set are reported in Supporting Information Table II.

Testing performance on high-resolution structures: The HRset

To test the CRK final parameters on an independent high-quality dataset, we constructed the above-mentioned HRset of interfaces from high-resolution crystal struc-

Table II

CRK Analysis Results of the $FimD_N$-FimC-$FimH_P$ Ternary Complex

Interface	Area (Å^2)	CA (%)	N	CRK	wCRK
$FimD_{N,\ red}$-FimH (Fig. 1)					
$FimD_N$-$FimH_P$	486	82	6	0.92	1.29
$FimD_{N,\ green}$-FimC (Fig. 1)					
$FimD_N$-FimC	1,078	95	8	0.25	0.32

The assembly choice featuring the smaller $FimD_N$-$FimH_P$ interface of Figure 1 ($FimD_{N,\ green}$-FimC-$FimH_P$) is the correct one, whereas the larger $FimD_{N,\ red}$-$FimH_P$ interface is a crystal contact (PDB entry 1ZE3).

tures, containing 105 crystal contacts and 32 biologically relevant interfaces. Of 19 calculated biological interfaces, 15 were correctly assigned (accuracy: 78.9%, recall: 46.9%). For crystal contacts, 52 were calculated and 48 correctly assigned (accuracy: 92.3%, monomer recall: 45.7%). The overall accuracy for the HRset is 88.7% with a recall of 46.0%. All results for the HRset are reported in Supporting Information Table III.

Comparing CRK with DiMoVo and PISA: Dey dimer dataset and HRset monomer subset

To compare the accuracy of CRK with DiMoVo[10] and PISA,[7] we used the monomeric part of the HRset and the main set of homodimeric proteins from Dey *et al.*[20] Entries that exhibited more than 1% of Ramachandran outliers in a MolProbity[22] analysis were removed from the datasets, as were membrane proteins and entries for which the oligomeric state could not be satisfyingly confirmed. For the Dey dimer dataset, CRK was able to assign 80 of the 125 interfaces, 56 of them correctly [accuracy 70.0% (DiMoVo: 68.8%, PISA: 94.0%), recall 44.8% (DiMoVo: 68.8%, PISA 88.0%)]. For the monomer subset of HRset, 52 of 105 interfaces were calculated and 48 were correctly assigned as crystal contacts [accuracy 92.3% (DiMoVo: 95.1%, PISA: 97.1%), recall 45.7% (DiMoVo: 92.4%, PISA 94.3%)].

The results for CRK, DiMoVo, and PISA, as well as more detailed statistics are reported in Supporting Information Table IV.

Exploring applications in structural biology

CRK was also applied to the crystallographic problem of the $FimD_N$-FimC-$FimH_P$ complex described in the Introduction section. The method provided an unambiguous answer, showing that the putative $FimD_N$-$FimH_P$ interface (red and yellow in Fig. 1) is a crystal contact, whereas the $FimD_N$-FimC interface displayed in green and cyan exhibits a clear signal for biological relevance, as shown by Nishiyama *et al.* (Table II).[1]

A very interesting test case and a "natural control" for CRK is represented by the structure of cypovirus polyhedrin,[24] a protein that forms extremely stable, micron-sized crystals to shield embedded viral particles. As polyhedrin crystals (e.g., PDB entry 2OH5) are a result of evolution, their crystal contacts are expected to be detected as biologically significant; indeed, the two main crystal contacts that form the lattice are categorized by CRK as biologically significant (Table III). As there are very few polyhedrin sequences available, the redundancy reduction procedure was deactivated (relaxing the BLAST identity threshold from 60% to 50% as for the Elcock and McCammon dataset would not help).

Table III

CRK Analysis for the Two Main Crystal Contacts Forming the Lattice of Polyhedrin Crystals

Interface	Area (Å^2)	CA (%)	N	CRK	wCRK
2OH5_1	1,812	95	13	0.57	0.62
2OH5_2	1,232	95	7	0.36	0.28

Identity plateau 50%, no redundancy reduction.

Another example of the potential of the CRK method comes from a differential analysis of two subsequent versions of the *S. typhimurium* MsbA ABC transporter structure in complex with ADP·vanadate plus lipopolysaccharide and with AMPPNP, respectively.[25,26] The former, incorrect structure (PDB entry 1Z2R) was retracted by the authors. Subsequently, the latter, correct structure was published with PDB code 3B60. MsbA is a functional dimer with a large interface, which is correctly recognized as biologically relevant in 3B60, but classified as a crystal contact in 1Z2R (Table IV).

Interestingly, in both cases, the interfaces are tightly packed (as denoted by the high core boundaries of 95% for both 3B60 and 1Z2R), but the incorrect positioning of residues in 1Z2R leads to very high CRK and wCRK values. This example highlights the possibility of using CRK for structure validation purposes.

The powerful PISA method identifies and classifies interfaces based on thermodynamic considerations, while CRK acts as an indicator for the selection pressure on the interface. Thus, the two approaches can be used together to assess difficult cases. An interesting example is represented by the decameric assembly in a new crystal form of BPTI by Lubkowski and Wlodawer.[27] The authors reported that no biological relevance for BPTI decamers was known or postulated. However, their finding

Table IV

Comparison of Two Structures of the ABC Transporter MsbA

Interface	Area (Å^2)	CA (%)	N	CRK	wCRK
1Z2R_1	2,842	95	9	4.14	4.49
1Z2R_2	2,531	95	9	3.80	4.36
3B60_1	6,945	95	45	0.57	0.70
3B60_2	6,923	95	48	0.56	0.71

1Z2R was made obsolete as it was incorrect. For 1Z2R, the secondary parameter "PISA identity threshold" (default 90%) had to be set as 89%.

Table V
CRK Analysis of the BPTI Decamer (PDB Entry 2HEX)

Interface	Area (Å^2)	CA (%)	N	CRK	wCRK
2HEX_1	569	95	6	0.46	0.42
2HEX_2	562	93	6	0.50	0.43
2HEX_3	540	91	6	0.51	0.43

Identity plateau 50%.

prompted further studies on the oligomeric state of BPTI in solution. Two articles, one describing magnetic relaxation dispersion experiments[28] and the other based on small-angle X-ray scattering[29] report that BPTI exists as a stable decameric species under a wide range of conditions. Both articles discuss the possible biological relevance of such decamers and come to diverging conclusions: while Gottschalk *et al.* proposed that the decamers exist as such in mast cell granules because of macromolecular crowding effect, Hamiaux *et al.* pointed out that they are unlikely to exist in the pancreatic secretion conditions. Analysis of the decameric structure (PDB entry 2HEX) with PISA categorizes its assembly as stable. Notably, a CRK analysis indicates that the key interfaces constituting the decamer are biologically significant (Table V).

In summary, CRK is a novel indicator to distinguish biologically relevant interfaces from crystal contacts, which rests on a different principle when compared with the existing tools. CRK is also potentially useful in the field of structure validation.

The above described results and examples show that the method yields accurate assignments: the overall accuracy of CRK is 84% and its accuracy for the DiMoVo set is 86% (in comparison, the DiMoVo program achieved 93% accuracy[10] with cutoff 0.5).

In the two datasets used for benchmarking, the accuracy of CRK was comparable with that of DiMoVo but substantially lower than that of PISA in the assignment of dimers (Dey dataset). Ample room remains for further improvement of the accuracy of CRK: this will involve, for instance, a more sophisticated sequence selection approach for the multiple alignment, or a flexible identity cutoff for sequence inclusion, optimized in each case to include the maximal possible amount of sequences while still avoiding K_s saturation.

CRK recall varied widely from dataset to dataset and was low in some instances: the key factor affecting recall is the availability of enough coding sequences meeting the selection criteria for a given protein or protein complex to be analyzed. This limiting factor is likely to become less important with the growth of nucleotide and protein databases, fuelled by the huge amount of genome sequencing projects being carried out worldwide on hundreds of species. Indeed, during the development phase, we could observe an increase in recall when switching to a new release of UniProt, and we expect that CRK recall will continue "growing" in the future.

ACKNOWLEDGMENTS

The authors thank José M. Duarte for scientific discussion and for carefully reading this article.

REFERENCES

1. Nishiyama M, Horst R, Eidam O, Herrmann T, Ignatov O, Vetsch M, Bettendorff P, Jelesarov I, Grutter MG, Wuthrich K, Glockshuber R, Capitani G. Structural basis of chaperone-subunit complex recognition by the type 1 pilus assembly platform FimD. EMBO J 2005;24:2075–2086.
2. Janin J. Specific versus non-specific contacts in protein crystals. Nat Struct Mol Biol 1997;4:973–974.
3. Bahadur RP, Chakrabarti P, Rodier F, Janin J. A dissection of specific and non-specific protein-protein interfaces. J Mol Biol 2004; 336:943–955.
4. Elcock AH, McCammon JA. Identification of protein oligomerization states by analysis of interface conservation. Proc Natl Acad Sci USA 2001;98:2990–2994.
5. Guharoy M, Chakrabarti P. Conservation and relative importance of residues across protein-protein interfaces. Proc Natl Acad Sci USA 2005;102:15447–15452.
6. Krissinel E, Henrick K. Detection of protein assemblies in crystals. In: Berthold MR, Glen R, Diederichs K, Kohlbacher O, Fischer I, editors. Computational life sciences, proceedings, Vol. 3695. Lecture notes in computer science. Berlin: Springer-Verlag; 2005. pp 163–174.
7. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. J Mol Biol 2007;372:774–797.
8. Ponstingl H, Henrick K, Thornton JM. Discriminating between homodimeric and monomeric proteins in the crystalline state. Proteins 2000;41:47–57.
9. Zhu H, Domingues F, Sommer I, Lengauer T. NOXclass: prediction of protein-protein interaction types. BMC Bioinformatics 2006;7:27.
10. Bernauer J, Bahadur RP, Rodier F, Janin J, Poupon A. DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein-protein interactions. Bioinformatics 2008;24:652–658.
11. Kobe B, Guncar G, Buchholz R, Huber T, Maco B, Cowieson N, Martin JL, Marfori M, Forwood JK. Crystallography and protein-protein interactions: biological interfaces and crystal contacts. Biochem Soc Trans 2008;36 (Part 6):1438–1441.
12. Stern A, Doron-Faigenboim A, Erez E, Martz E, Bacharach E, Pupko T. Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. Nucleic Acids Res 2007;35 (Web Server Issue):W506–W511.
13. Doron-Faigenboim A, Stern A, Mayrose I, Bacharach E, Pupko T. Selecton: a server for detecting evolutionary forces at a single amino-acid site. Bioinformatics 2005;21:2101–2103.
14. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
15. Notredame C, Higgins DG, Heringa J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 2000;302: 205–217.
16. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
17. Uniprot_Consortium. The Universal Protein Resource (UniProt) 2009. Nucleic Acids Res 2009;37 (Database Issue):D169–D174.
18. Yang Z, Nielsen R, Goldman N, Pedersen A-M. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 2000;155:431–449.
19. Percudani R, Peracchi A. A genomic overview of pyridoxal-phosphate-dependent enzymes. EMBO Rep 2003;4:850–854.

20. Dey S, Pal A, Chakrabarti P, Janin J. The subunit interfaces of weakly associated homodimeric proteins. J Mol Biol 2010;398:146–160.
21. Eliot AC, Kirsch JF. Pyridoxal phosphate enzymes: mechanistic, structural, and evolutionary considerations. Annu Rev Biochem 2004;73:383–415.
22. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB, III, Snoeyink J, Richardson JS, Richardson DC. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res 2007;35 (Web Server Issue):W375–W383.
23. DeLano WL. The PyMOL molecular graphics system. 2002. DeLano Scientific, San Carlos, CA, USA. http://www.pymol.org.
24. Coulibaly F, Chiu E, Ikeda K, Gutmann S, Haebel PW, Schulze-Briese C, Mori H, Metcalf P. The molecular organization of cypovirus polyhedra. Nature 2007;446:97–101.
25. Reyes CL, Chang G. Structure of the ABC transporter MsbA in complex with ADP.vanadate and lipopolysaccharide. Science 2005;308:1028–1031.
26. Ward A, Reyes CL, Yu J, Roth CB, Chang G. Flexibility in the ABC transporter MsbA: alternating access with a twist. Proc Natl Acad Sci USA 2007;104:19005–19010.
27. Lubkowski J, Wlodawer A. Decamers observed in the crystals of bovine pancreatic trypsin inhibitor. Acta Crystallogr D Biol Crystallogr 1999;55 (Part 1):335–337.
28. Gottschalk M, Venu K, Halle B. Protein self-association in solution: the bovine pancreatic trypsin inhibitor decamer. Biophys J 2003;84: 3941–3958.
29. Hamiaux C, Perez J, Prange T, Veesler S, Ries-Kautt M, Vachette P. The BPTI decamer observed in acidic pH crystal forms pre-exists as a stable species in solution. J Mol Biol 2000;297:697–712.