BIOINFORMATICS ORIGINAL PAPER
Vol. 24 no. 5 2008, pages 652–658
doi:10.1093/bioinformatics/btn022

Structural bioinformatics

DiMoVo: a Voronoi tessellation-based method for discriminating crystallographic and biological protein–protein interactions

Julie Bernauer[1,2], Ranjit Prasad Bahadur[2], Francis Rodier[3], Joël Janin[2] and Anne Poupon[2,*]

[1]Department of Structural Biology, Fairchild Science Building, Stanford University School of Medicine, Stanford, CA 94305-5126, USA, [2]Yeast Structural Genomics, IBBMC UMR 8619, bâtiment 430, Université Paris-Sud, 91405 Orsay and [3]LEBS UPR 9063, CNRS, 91191 Gif-sur-Yvette, France

Received on October 30, 2007; revised and accepted on January 10, 2008

Advance Access publication January 19, 2008

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Knowledge of the oligomeric state of a protein is often essential for understanding its function and mechanism. Within a protein crystal, each protein monomer is in contact with many others, forming many small interfaces and a few larger ones that are biologically significant if the protein is a homodimer in solution, but not if the protein is monomeric. Telling such 'crystal dimers' from real ones remains a difficult task.

Results: It has already been demonstrated that the interfaces of native and non-native protein–protein complexes can be distinguished using a combination of parameters computed with a method on the Voronoi tessellation. We show in this article that the same parameters highlight significant differences between the interfaces of biological and crystal dimers. Using these parameters as descriptors in machine learning methods leads to accurate classification of specific and non-specific protein–protein interfaces.

Availability: Software is available at http://fifi.ibbmc.u-psud.fr/DiMoVo

Contact: anne@rezo.net

1 INTRODUCTION

Proteins fulfill almost all the 'active' roles in life: enzymes, antibodies, receptors, etc. For most proteins, biological and biochemical function can be accomplished only through the association of two or more macromolecules. Among the different assemblies that two or more polypeptide chains (monomers) can form, homo-oligomers contain multiple copies of the same monomers, whereas hetero-oligomers and complexes associate different proteins.

The structure of proteins and protein assemblies is experimentally accessible through the interpretation of images obtained by the diffraction of X-rays by a protein crystal. In a crystal, each monomer makes contacts with many others, identical or not. These contacts are non-specific and do not exist *in vivo*, but they are of the same physico-chemical nature as the specific contacts that stabilize complexes and oligomeric proteins. It has been shown that most of the pairwise interfaces created by the crystal packing are smaller than those of complexes and oligomeric proteins (Bahadur *et al*., 2003, 2004; Carugo and Argos, 1997; Dasgupta *et al*., 1997; Janin, 1997). However, some are comparable in size to specific interfaces, and a pair of protein molecules that form a large packing interface in the crystal may be mistaken for a biological homodimer if it has 2-fold symmetry. We call such pairs 'crystal dimers'.

Some dimeric proteins (either homodimers or heterodimers) are called 'obligate' dimers. These are proteins than can exist only associated to each other, and are never found in the monomeric state [see Nooren and Thornton (2003) for a review]. This is the case for the bacteriophage P22 Arc repressor [PDB code: 1ARR (Bonvin *et al*., 1994)]. The repressor consists in two identical chains that fold together, and both chains participate in the hydrophobic core (Milla *et al*., 1995). Other dimers are called non-obligates, and can be found as monomers. In these cases, complexation often regulates the activity of the protein. This is the case of sperm lysine (PDB code: 1LYN), which is active as a monomer and inactive as a dimer (Shaw *et al*., 1995).

Understanding the function of a protein often requires the knowledge of its oligomeric state. Whereas this should be done by experiment in solution, one may attempt to derive the oligomeric state from the crystal structure when it is available. This has proved surprisingly difficult. A recent study uses graph theory to find all the possible assemblies in the crystal, and then computes, for each interface, a quantity related to the free energy change upon dissociation. The PISA server reports the results (Krissinel and Henrick, 2007). Another server, PITA (Ponstingl *et al*., 2003) scores crystallographic contacts by their contact size and chemical complementarity. NOXClass (Zhu *et al*., 2006) uses descriptors of the interface, such as its size, its composition, or the chemical complementarity of the contacting residues, as parameters for a support vector machine procedure. Block *et al*. (2006) used feature selection methods, coupled with four different machine learning methods. Liu *et al*. (2006) use combinations of different physico-chemical and geometrical parameters.

*To whom correspondence should be addressed.

When applied to test sets of specific interfaces in biological dimers and of non-specific interfaces in crystal dimers, these methods discriminate between the two types with success rates near 92% (Mintseris and Weng, 2003; Ponstingl *et al.*, 2003; Zhu *et al.*, 2006). However, the size of the interface is a predominant parameter in all cases, and the presence of many small interfaces in the non-specific set greatly helps. Indeed, the success rate drops drastically when the non-specific set contains crystal dimers with interfaces similar in size to those in the specific set.

We present here a method based on the Voronoi tessellation and a coarse-grained modelling of the protein structure, and use it to discriminate between the interfaces of biological homodimers and packing interfaces of a similar size formed by monomeric proteins in crystals. This new method, called DiMoVo (DIscrimination between Multimers and MOnomers by VOronoi tessellation), achieves a high accuracy even on crystal dimers that have large interfaces.

2 METHODS

2.1 Modeling the protein structure

Although the three dimensional structure of proteins is considered unique, atoms on the surface are weakly constrained, and many atomic movements happen when two proteins interact. As they are very difficult to predict, a low-resolution, simplified model, is often more appropriate than a detailed atomic structure when modeling protein–protein interaction. We represent protein structures by spheres representing amino acid residues and build a Voronoi tessellation to calculate a set of parameters that have proven useful in combination with machine learning algorithms (Bernauer *et al.*, 2005, 2007; Poupon, 2004).

Residue-based Voronoi tessellations were built with a previously described procedure (Bernauer *et al.*, 2007) that uses the Computational Geometry Algorithms Library (CGAL) (Boissonnat *et al.*, 1999) and was optimized to take less than one second for each complex.

2.2 Parameters

As described in (Bernauer *et al.*, 2005), the parameters are:

- the interface area (1 parameter)
- the number of core interface residues (1 parameter)
- their Voronoi volumes (20 parameters)
- the frequency of each residue type at the interface (20 parameters)
- the frequency of the pairs of residues in contact (210 parameters)
- the distances between their geometric center (210 parameters)

The latter two categories should comprise 210 parameters each, a number reduced to 21 by grouping residues in six bins: Hydrophobic (VILM), Aromatic (FYW), Positive (HKR), Negative (DE), Polar non-charged (NQ), Small (AGSTCP).

In addition, we tested some of the parameters described in (Bahadur *et al.*, 2004):

- Residue Propensity score (RP), for a residue of type *i*, this propensity is computed as the logarithm of the ratio between the area fraction contributed to the protein–protein interface and the area fraction contributed to the protein–solvent interface by residues of type *i*.
- Global Density index (GD): the mean number of atoms per unit area in an ellipse fitting the interface.
- Local Density index (LD): the mean number of neighbors of an interface atom.

2.3 Data sets

2.3.1 Training set The data sets used for developing the method are those of (Bahadur *et al.*, 2003) for the biological dimers and (Bahadur *et al.*, 2004) for the crystal dimers. The proteins of these sets were selected manually from the Protein Data Bank, and their status of dimer or monomer in solution was checked with the biochemical literature. In addition, we required that the sequence of the crystallized fragment has to be the one used for multimeric studies. Indeed, experimental results showing that the full length protein forms a stable dimer cannot guaranty that a fragment will also form a stable dimer. The monomer set contains 178 crystal dimers selected to have an interface area greater than 800 Å^2. The biological dimer set contains 113 biological homodimers.

2.3.2 Test sets We used two other data sets to compare our method with three other: NOXClass (Zhu *et al.*, 2006), PISA (Krissinel and Henrick, 2005) and PITA (Ponstingl *et al.*, 2003).

- The Zhu dataset (Zhu *et al.*, 2006). This dataset, compiled from previously published sets (Bradford and Westhead, 2005; Neuvirth *et al.*, 2004), contains 'obligate', 'non-obligate' interactions and crystal contacts. For this study, 'obligates' were equated with homodimers and 'non-obligates' were not considered since these were only heterodimers. This dataset has been used to train the NOXClass method.
- The Ponstingl dataset (Ponstingl *et al.*, 2003). This dataset contains crystallographic dimers, biological dimers, and biological multimers of larger order. For this study, only the 'monomers' and 'dimers' categories were retained. This dataset has been used to train PISA (Krissinel and Henrick, 2007) and PITA (Ponstingl *et al.*, 2003).

It should also be noted that a few examples from these two datasets were removed for different reasons. We removed items in these datasets when: (i) the two polypeptide chains were not identical, (ii) a co-factor or metallic ion was present at the interface, (iii) absence of a symmetry matrix in file containing only one peptide chain, (iv) one of the four tested methods fails to evaluate the example, (v) the polypeptide chain was a fragment of the protein for which the quaternary structure has been experimentally established or (vi) the quaternary structure had been derived from homologous proteins and not experimentally established.

2.4 Learning

The values of the 84 Voronoi-derived parameters and the indexes RP, GD and LD measured on the Bahadur datasets of monomers and homodimers were used as input for a Support Vector Machine (SVM) procedure (Cristiani and Shawe-Taylor, 2000; Schölhopf, 1997) using a Radial Basis Function (RBF) kernel. As in Bernauer (2006), missing data were replaced by the median value of the corresponding parameter on the whole dataset.

To ensure reliable statistics, and avoid over-fitting, learning was done in leave-one-out with a 5-fold cross-validation procedure.

The R libSVM package (Chang and Lin, 2001) was used to perform support vector machines experiments. The optimization of *C* and γ was done by a grid search (Hsu *et al.*, 2003) using the 'tune' function of libSVM. The ROC curves analyses were performed using the packages 'ROCR' (Sing *et al.*, 2005) and 'verification'.

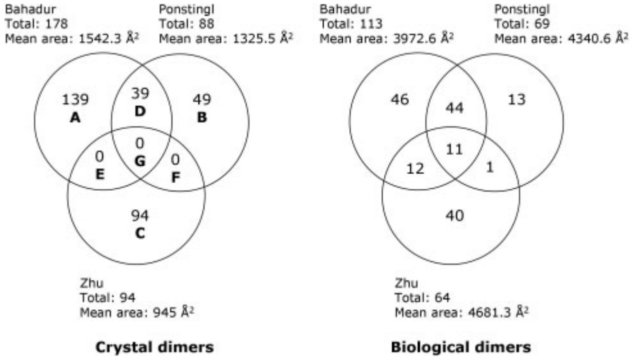

Fig. 1. learning datasets. For crystal and biological dimers, region A contains PDB codes found only in Bahadur dataset, region B those found only in the Ponstingl dataset, and region C those found only in Zhu datasets. Regions D, E and F are those examples found in two of the datasets, but not in the third one. Region G contains the examples common to the three datasets. For each dataset, and for both crystal and biological dimers, the total number of examples and the mean interface area are given.

2.5 Nested tests

We performed nested tests to evaluate the contribution of each parameter and selected the best performing subset. Learning was repeated after eliminating each of the 84 parameters in turn. The parameter whose elimination leads to the highest accuracy is excluded. Each of the 83 remaining parameters is eliminated in turn, and again, the parameter whose elimination leads to the highest accuracy is excluded. The procedure was repeated until only two parameters remained. Each learning procedure was done through leave-one-out.

2.6 Accuracy and recall evaluation

The three datasets (Bahadur, Zhu and Ponstingl) have important overlaps (Fig. 1), comparing the accuracies and recalls (on both monomers and dimers) requires that each method is evaluated only on proteins not belonging to their training dataset.

For each method, a test set was constructed using all the proteins that were not present in the tested method's training set, and had no more than 30% sequence identity with proteins of the training set. Moreover, the internal redundancy of each test set was filtered out so that two proteins in the set don't share more than 30% identity. The test sets are available on the website.

As the authors of PITA estimate that scores above 70–80 indicate specific interfaces (Ponstingl *et al.*, 2003), (i) scores below 70 were taken to identify crystal dimers, (ii) scores above biological dimers, (iii) interfaces with scores between 70 and 80 are not assigned.

The overall accuracy is the number of correctly assigned examples divided by the number of assigned examples. The crystal dimer (resp. biological dimer) accuracy is the number of correctly assigned crystal dimers (resp. biological dimers) divided by the total number of assigned crystal dimers (resp. biological dimers). The crystal dimer (resp. biological dimer) recall is the number of correctly assigned crystal dimers (resp. biological dimers) divided by the total number of crystal dimers (resp. biological dimers), assigned or not.

3 RESULTS

3.1 Values of the parameters

The area of interface (computed as the sum of the areas of Voronoi facets shared by cells from the two subunits), and the number of interface residues are on average much smaller for crystal dimers than for biological dimers (Fig. 2A and B). These two parameters are highly correlated.

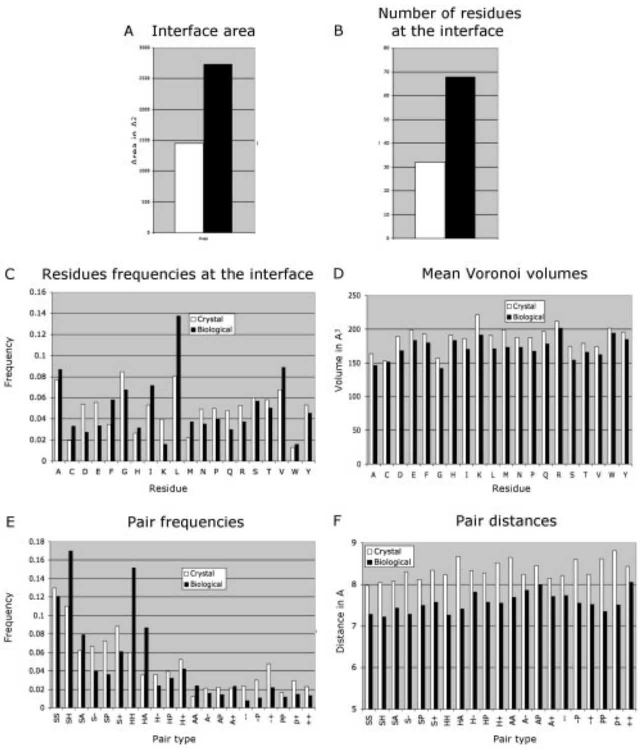

Fig. 2. Values of the parameters for crystal and biological dimers. The values of the 84 parameters used for learning are shown in black bars for biological dimers and white bars for crystal dimers. (**A**) Voronoi area of the interface. (**B**) Number of residues participating in the interface. (**C**) Frequencies of each of the 20 residues at the interface. (**D**) Mean volume of the Voronoi cell for each of the 20 residues at the interface. (**E**) Frequency of each type of pair at the interface. (**F**) Mean distance between the geometric centers in a pair at the interface for each type of pair. In E and F, S: small residue (AGSTCP), H: hydrophobic (VILM), A: aromatic (FYW), −: negatively charged (DE), +: positively charged (NQ), P: polar (HKR).

Figure 2C shows the frequency of the 20 amino acids at the interfaces. As previously described, the biological interfaces are enriched in hydrophobic residues and depleted in polar and charged residues relative to crystal packing interfaces, and also to the solvent-accessible surface (Bahadur *et al.*, 2004). Small residues have similar frequencies in both types of interfaces, except cysteine, which is more frequent in biological interfaces.

Our previous study of protein–protein complexes (Bernauer *et al.*, 2007) indicates that the mean volumes of the Voronoi cells of interface residues are important parameters. Figure 2D shows some interesting variation. The volume occupied by leucines, which is somewhat larger at biological interfaces than in the protein core, is even larger at crystal packing interfaces.

The frequencies of residue pairs in contact at the interface also show interesting differences (Fig. 2E). The hydrophobic–hydrophobic, hydrophobic-small and hydrophobic–aromatic pairs are much more frequent in biological interfaces. This is partly due to the enrichment of biological interfaces in hydrophobic residues, but also to their better physico-chemical complementary compared to crystal packing interfaces.

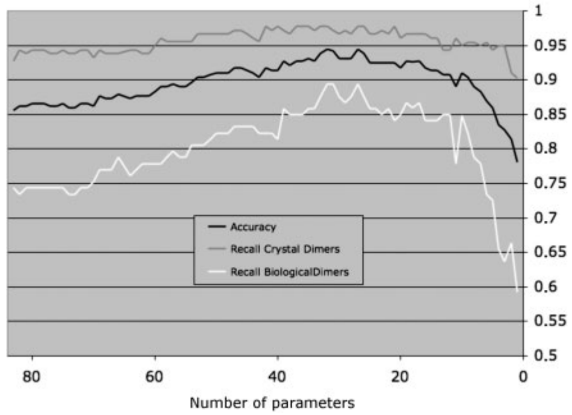

Fig. 3. Accuracy (white line), crystal dimer recall (grey line) and biological dimer recall (black line) as a function of the number of parameters. At each step, the accuracy with all parameters but one is evaluated. The parameter giving the highest accuracy (consequently the parameter that is less informative) is eliminated for the next step.

3.2 Parameter selection

In order to evaluate the relative importance of the 84 parameters, we performed nested tests. The most important thing we learned during these tests is that the method is more accurate if not all parameters are used (Fig. 3). From these tests, we concluded that 27 parameters should be retained (see Table 1). Using these 27 parameters leads to an accuracy of 0.95, with a crystal dimer recall of 0.98 and a biological dimer recall of 0.89.

As expected, the most important parameter is the interface area (Table 1): alone, it yields an accuracy of 0.78 with a crystal dimer recall of 0.90 and a biological dimer recall of 0.59. More surprising is the fact that both area of the interface and number of residues at the interface, although largely correlated, are very important (the number of residues at the interface is ranked 9).

A well represented type of parameter in the retained set is the volume of the Voronoi cell. Frequencies are also well represented in the selected set. In both cases, the importance of these parameters cannot be correlated to their repartitions between biological and crystal packing interfaces. Pair frequencies appear less often, however, the frequency of small-polar pairs is ranked 4. Since this type of pair is less frequent in biological interfaces, its ranking shows that the presence of this type of pair in a biological interface

Table 1. Rank of the parameters in nested tests

Type	AA	Rank	Type	AA	Rank	Type	AA	Rank
Area		1	Volume	I	29	Pair freq	++	57
Volume	L	2	Volume	M	30	Pair dist	HP	58
Volume	S	3	Pair freq	SA	31	Pair dist	PP	59
Pair freq	SP	4	Volume	N	32	Frequency	M	60
Frequency	D	5	Volume	P	33	Volume	Y	61
Frequency	F	6	Pair freq	SS	34	Pair dist	++	62
Frequency	Y	7	Frequency	Q	35	Pair freq	A+	63
Frequency	L	8	Frequency	A	36	Pair freq	AP	64
Nb Res		9	Pair freq	H+	37	Pair freq	HA	65
Volume	C	10	Frequency	R	38	Pair dist	A+	66
Frequency	G	11	Volume	W	39	Pair dist	−P	67
Frequency	K	12	Volume	D	40	Pair dist	S+	68
Pair freq	AA	13	Pair freq	H−	41	Pair freq	S−	69
Volume	V	14	Volume	A	42	Pair freq	PP	70
Frequency	S	15	Volume	G	43	Frequency	H	71
Pair dist	S−	16	Frequency	P	44	Pair dist	A−	72
Volume	K	17	Frequency	C	45	Frequency	E	73
Pair freq	HP	18	Pair dist	SS	46	Volume	R	74
Pair dist	HH	19	Volume	Q	47	Pair dist	HA	75
Frequency	I	20	Pair freq	SH	48	Pair dist	SP	76
Pair dist	SH	21	Pair freq	−−	49	Pair freq	P+	77
Frequency	N	22	Frequency	W	50	Pair dist	AA	78
Volume	T	23	Volume	F	51	Pair dist	SA	79
Pair freq	−P	24	Pair freq	S−	52	Pair dist	−+	80
Pair dist	H+	25	Pair freq	A−	53	Pair dist	P+	81
Pair dist	AP	26	Volume	H	54	Frequency	T	82
Pair dist	−−	27	Pair freq	−+	55	Pair freq	HH	83
Volume	E	28	Pair dist	H−	56	Frequency	V	84

Pair dist: pair distance, Pair freq: pair frequency, Nb res: number of residues at the interface, Area: area of the interface. A: Aromatic, H: Hydrophobic, S: Small, P: Polar, +: positively charged, −: negatively charged. Parameters on grey background are those retained by the nested tests analysis.

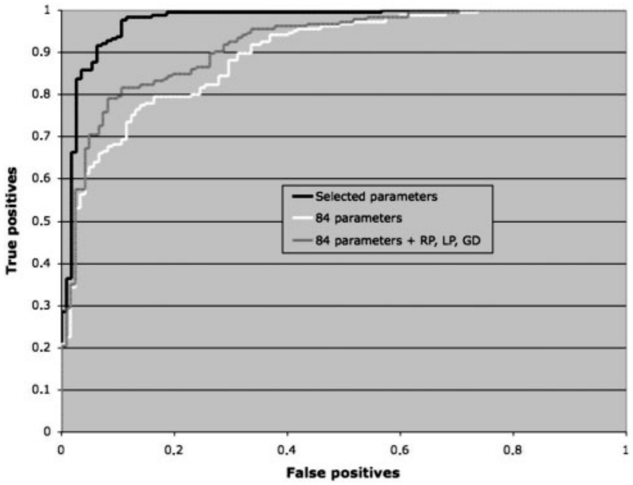

Fig. 4. ROC curves for three different sets of parameters: the selected set of 27 parameters (see text), the original set of 84 parameters, and the 84 low-resolution parameters with the three high-resolution parameters RP, LP and GD (see text).

is unfavorable. Similarly, pair distances, which were thought to have a very weak contribution, do appear in the retained set.

To summarize, these result show that the factors that favor biological interfaces are:

- a larger interface area,
- more interface residues,
- a large proportion of F, I and L, a small proportion of D, G, K, N, S and Y,
- small volumes for C, K, L, S, T and V,
- a large number of aromatic–aromatic pairs and a small number of small-polar, hydrophobic–polar and negative-polar pairs,
- small distances in small-negative, hydrophobic–hydrophobic, hydrophobic-positive, aromatic–polar and negative–negative pairs.

3.3 Learning performance

Performance of learning has been evaluated by the Area Under the ROC Curve (AUC). Receiver Operating Characteristics (ROC) curves, issued from signal processing, represent the trade-off between true and false positive rates when interpreting hypotheses. The ideal hypothesis, which generates no false positive and 100% of the true positives, has a ROC curve that is a step function and an AUC of 1. A random selection yielding true and false positives in equivalent numbers has an AUC of 0.5.

Figure 4 shows three ROC curves: learning with 84 parameters, learning with 84 parameters and high-resolution geometric parameters RP, LP and GD, and learning with the selected set of 27 parameters. The curves show that whereas adding the RP, LP and GD parameters has only a minor effect on accuracy (the AUC increases from 0.91 to 0.92), a reduced set of parameters performs much better, with an AUC of 0.97.

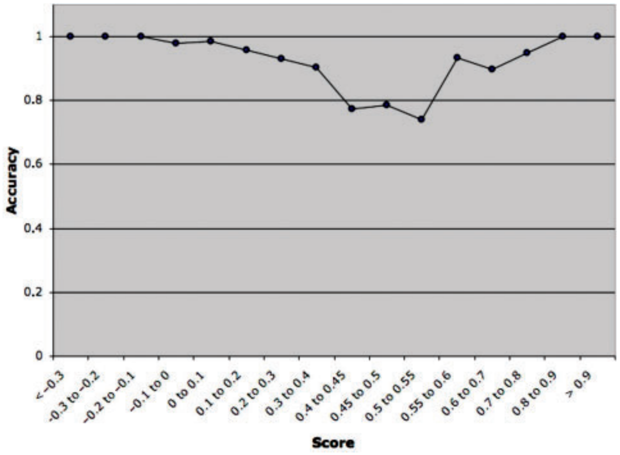

Fig. 5. Accuracy of DiMoVo as a function of score.

3.4 Reliability of predictions

Because the SVM, for each submitted example, returns a score, we are able to give a reliability rate for a prediction, depending on the score obtained (Fig. 5).

As expected, since the SVM aims at giving to a crystal dimer a score close to 0, and to a biological dimer a score close to 1, the accuracy of predictions with scores lower than 0.3, or higher than 0.8 is very high. Although not very low, the accuracy around 0.5 is not as good (0.73 accuracy for scores between 0.5 and 0.55, and 0.79 accuracy for scores between 0.45 and 0.5). Consequently, we tested three different options:

- threshold 0.5: if the score is lower than 0.5, we predict a crystal dimer, if the score is higher than 0.5, we predict a biological dimer
- 0.45–0.55: if the score is lower than 0.45, we predict a crystal dimer, if the score is higher than 0.55, we predict a biological dimer, if the score is between 0.45 and 0.55 we don't predict.
- 0.4–0.6: if the score is lower than 0.4, we predict a crystal dimer, if the score is higher than 0.6, we predict a biological dimer, if the score is between 0.4 and 0.6 we don't predict.

Results obtained in leave-one-out on the learning set are given in Table 2. DiMoVo in its simplest flavor (with a threshold at 0.5) is already performing well with an accuracy of 0.95. The important difference between crystal and biological dimer recalls (0.98 and 0.89, respectively) is largely due to the fact that there are more crystal than biological dimers in the learning set. However, randomly removing crystal dimers from the set, which is the usual methodology in this type of method, led to significantly lower accuracies. This could have been due to over-fitting. However, the accuracy and recalls obtained on the two other datasets show that we are not in this situation. This is also confirmed by the number of vectors (171 ± 12), which is significantly lower than the number of structures used in the learning set.

When considering that structures with scores between 0.45 and 0.55 (in DiMoVo 0.45–0.55), or scores between 0.4–0.6

Table 2. Accuracy, crystal dimer recall and biological dimer recall for three different flavors of DiMoVo, obtained on the learning set, using a leave-one-out procedure

Method	Accuracy	Crystal dimers	Biological dimers
DiMoVo 0.5	0.95	0.98	0.89
DiMoVo 0.45–0.55	0.96	0.95	0.8
DiMoVo 0.4–0.6	0.97	0.92	0.73

DiMoVo 0.5: if score is lower than 0.5, a crystal dimer is predicted, else a biological dimer is predicted. DiMoVo 0.45–0.55: if score is lower than 0.45, a crystal dimer is predicted, if score is higher than 0.55 a biological dimer is predicted, else no prediction. DiMoVo 0.4–0.6: if score is lower than 0.4, a crystal dimer is predicted, if score is higher than 0.6 a biological dimer is predicted, else no prediction is made.

(in DiMoVo 0.4–0.6) cannot be assigned to crystal or biological dimers, the accuracy is improved, but the recalls are lower.

3.5 Comparison with other existing methods

In order to assess the accuracy of our method, and to compare with already existing methods, we have used the four already cited methods: DiMoVo, PITA (Ponstingl *et al.*, 2003), PISA (Krissinel and Henrick, 2005) and NOXclass (Zhu *et al.*, 2006). We used the three datasets (Bahadur, Ponstingl and Zhu) for the tests. Each method was evaluated only on crystal and biological dimers that were not part of its own training set, and that have no more than 30% sequence identity with proteins of the learning set. Because of the large overlap between the different datasets, the number of examples used for the evaluation of each method is different (except for PITA and PISA which were trained on the same dataset).

The results are given Table 3. As can be seen, DiMoVo has both a better overall accuracy, but also better crystal dimer recall than the other three methods. PISA and NOXClass have a better biological dimer recall than DiMoVo, but this is counter-balanced by rather low crystal dimer recalls.

To further analyze predictions, and for comparison, we studied the reliability of predictions by the four methods as a function of the interface area. As can be seen in Figure 6, all methods show lesser accuracies for 'problematic' cases: those with interface areas from 1400 to 3000 Å^2. However, DiMoVo accuracies are never lower than 0.8 (areas between 2000 and 2500 Å^2, 0.83 for DiMoVo 0.4–0.6), whereas other methods show significantly lower values: 0.73 for PITA (areas between 2000 and 2500 Å^2), 0.5 for PISA (areas between 2500 and 3000 Å^2), and 0.57 for NOXClass (areas between 2000 and 2500 Å^2). The comparatively lower accuracies obtained with PITA and PISA are attenuated by the fact that these two methods are not only able to predict monomers and dimers, but also multimers. Moreover, PISA gives many very useful details concerning each possible interface. The larger number of possible predictions lowers the accuracy of one particular prediction, namely the discrimination between crystal and biological dimer.

Table 3. Comparison of the performances of 4 different methods: NOXClass (Zhu *et al.*, 2006), PISA (Krissinel and Henrick, 2005), PITA (Ponstingl *et al.*, 2003) and DiMoVo

Method	Crystal[a]	Biological[b]	Total	Acc	Recall C	Recall B
NOXClass	178	79	257	0.76	0.68	0.95
PISA	227	76	303	0.86	0.76	0.92
PITA	227	76	303	0.92	0.91	0.84
D. 0.5	137	33	170	0.93	0.95	0.84
D. 0.45–0.55	137	33	170	0.94	0.93	0.79
D. 0.4–0.6	137	33	170	0.97	0.91	0.7
D Zhu OC	106	75	181	0.90	0.97	0.79
D Zhu ONOC	106	137	243	0.88	0.91	0.85
D Ponstingl	89	69	158	0.83	0.92	0.71

[a] Number of crystal dimers.
[b] Number of biological dimers.

Each method is evaluated only on structures not part of the training set. Three different flavors of DiMoVo trained on Bahadur dataset are presented (D. 0.5, D. 0.45–0.55 and D. 0.4–0.6, see text). Results obtained with DiMoVo trained on: obligates and crystal dimers of Zhu dataset (D. Zhu OC); on obligates, non-obligates and crystal dimers of Zhu dataset (D. Zhu ONOC) and on dimers and monomers of Ponstingl dataset (D. Ponstingl), Recall C: recall crystal dimers. Recall B: recall biological dimers.

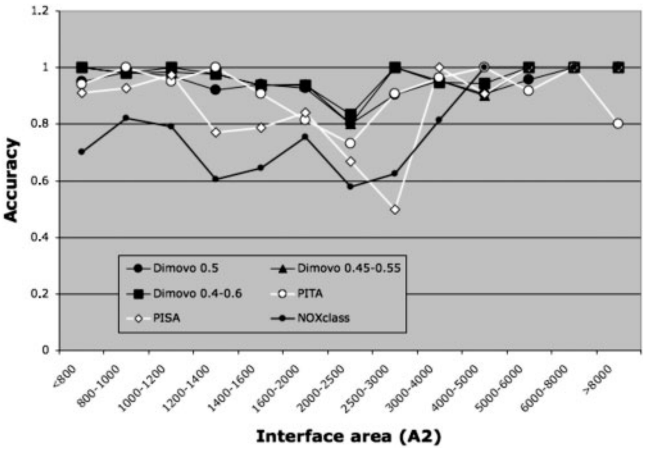

Fig. 6. Accuracy as a function of the interface area. The accuracies of crystal/biological assignment by PITA, PISA, NOXClass and the three different flavors of DiMoVo (see text) were estimated for examples with interface areas in different intervals.

Our improved performances for difficult cases are due to three facts. Firstly, our learning set contains more difficult cases, and especially more crystal dimers with large interface areas. To highlight this point, we have trained DiMoVo using (i) obligates and crystal dimers from Zhu dataset, (ii) obligates, non-obligates and crystal dimers from Zhu dataset and (iii) monomers and dimers form Ponstingl dataset. The results in Table 3 clearly show that these three learning sets lead to lower accuracies and recall. Secondly, and this is a direct consequence of the previous point, interface area, although a very important parameter, has a lower weight in our method as compared to other ones. Finally, we consider many more parameters than other methods, and use these to build a scoring function using a machine-learning procedure.

4 CONCLUSION

In this study, we set up an effective method to discriminate between crystal packing and biological interactions using a coarse-grained model for the structure. This method correctly discriminates between crystal and biological dimers for 97% of the tested cases (DiMoVo 0.4–0.6), with very good recalls for both types of dimers. DiMoVo has been shown to compare very favorably with existing methods, especially for difficult cases, namely those for which the area of the interface is around 2000 Å^2.

Interestingly, the initial set of 84 parameters had to be reduced to a subset of 27 parameters to obtain maximal accuracy. The parameters of this subset belong to all of the initial categories: volumes, frequencies, pair frequencies and pair distances. It has also shown that strong correlation of the two parameters was not a criterion for excluding one. Indeed, the number of residues constituting the interface, and the area of the interface, which are two strongly correlated parameters, are both present in the final subset, and excluding one decreases significantly the accuracy.

This work shows that the formalism used for the description of the protein structure is very powerful. Although only one point per residue is used, and the relative weights of the different types of residues are ignored, the obtained Voronoi tessellation allows computing well-discriminating descriptors. To try to improve further the discrimination between biological and crystal contact interfaces, we will try to use more complex mathematical tessellations, such as power diagrams. However, the first attempts we have made in this direction show that it is very difficult to adjust the weights assigned to each type of residue, since no simple relation can be established between these weights and the size or molecular weights of the residues. Moreover, if the parameters obtained show a larger gap between specific and non-specific interfaces, the standard deviations are also larger.

ACKNOWLEDGEMENT

J. Bernauer thanks the Association pour la Recherche contre le Cancer (ARC) for financial support.

Conflict of Interest: none declared.

REFERENCES

Bahadur,R.P. *et al.* (2003) Dissecting subunit interfaces in homodimeric proteins. *Proteins*, **53**, 708–719.

Bahadur,R.P. *et al.* (2004) A dissection of specific and non-specific protein-protein interfaces. *J. Mol. Biol.*, **336**, 943–955.

Bernauer,J. *et al.* (2007) A new protein–protein docking scoring function based on interface residue properties. *Bioinformatics*, **23**, 555–562.

Bernauer,J. *et al.* (2005) A docking analysis of the statistical physics of protein–protein recognition. *Phys. Biol.*, **2**, S17–S23.

Block,P. *et al.* (2006) Physicochemical descriptors to discriminate protein–protein interactions in permanent and transient complexes selected by means of machine learning algorithms. *Proteins*, **65**, 607–622.

Boissonnat,J.-D. *et al.* (1999) Programming with CGAL: The example of triangulations. *Symp. Comput. Geometry*, **1999**, 421–422.

Bonvin,A.M. *et al.* (1994) Nuclear magnetic resonance solution structure of the arc repressor using relaxation matrix calculations. *J. Mol. Biol.*, **236**, 328–341.

Bradford,J.R. and Westhead,D.R. (2005) Improved prediction of protein–protein binding sites using a support vector machines approach. *Bioinformatics*, **21**, 1487–1494.

Carugo,O. and Argos,P. (1997) Protein–protein crystal-packing contacts. *Protein Sci.*, **6**, 2261–2263.

Chang,C.-C. and Lin,C.-J. (2001) LIBSVM: a library for support vector machines. *Multiple Classifier Systems 6th International Workshop*, Seaside, CA, USA.

Cristiani,N. and Shawe-Taylor,J. (2000) *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge University Press, Cambridge.

Dasgupta,S. *et al.* (1997) Extent and nature of contacts between protein molecules in crystal lattices and between subunits of protein oligomers. *Proteins*, **28**, 494–514.

Hsu,C. *et al.* (2003) *A Practical Guide to Support Vector Classification*. Department of Computer Science and Information Engineering, National Taiwan University Editions, Taiwan.

Janin,J. (1997) Specific versus non-specific contacts in protein crystals. *Nat. Struct. Biol.*, **4**, 973–974.

Krissinel,E. and Henrick,K. (2007) Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.*, **372**, 774–797.

Liu,S. *et al.* (2006) A combinatorial score to distinguish biological and nonbiological protein–protein interfaces. *Proteins*, **64**, 68–78.

Milla,M.E. *et al.* (1995) P22 Arc repressor: transition state properties inferred from mutational effects on the rates of protein unfolding and refolding. *Biochemistry*, **34**, 13914–13919.

Mintseris,J. and Weng,Z. (2003) Atomic contact vectors in protein-protein recognition. *Proteins*, **53**, 629–639.

Neuvirth,H. *et al.* (2004) ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol.*, **338**, 181–199.

Nooren,I.M. and Thornton,J.M. (2003) Diversity of protein–protein interactions. *EMBO J.*, **22**, 3486–3492.

Ponstingl,H. *et al.* (2003) Automatic inference of protein quaternary structure from crystals. *J. Appl. Cryst.*, **36**, 1116–1122.

Poupon,A. (2004) Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. *Curr. Opin. Struct. Biol.*, **14**, 233–241.

Shaw,A. *et al.* (1995) Crystal structure and subunit dynamics of the abalone sperm lysin dimer: egg envelopes dissociate dimers, the monomer is the active species. *J. Cell. Biol.*, **130**, 1117–1125.

Schölhopf,B. (1997) *Support Vector Learning*. R. Oldenbourg Verlag, Munich.

Sing,T. *et al.* (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.

Zhu,H. *et al.* (2006) NOXclass: prediction of protein-protein interaction types. *BMC Bioinformatics*, **7**, 27.