

1

2 **PB-kPRED: knowledge-based prediction of protein backbone conformation using a**
3 **structural alphabet**

4 Iyanar Vetrivel^{1,5}, Swapnil Mahajan^{1,5}, Manoj Tyagi^{2,#}, Lionel Hoffmann¹, Yves-Henri Sanejouand¹,
5 Narayanaswamy Srinivasan³, Alexandre G. de Brevern⁴, Frédéric Cadet^{5,6}, Bernard Offmann^{1*}

7 ¹Université de Nantes, Unité Fonctionnalité et Ingénierie des Protéines (UFIP), UMR 6286 CNRS, UFR Sciences et
8 Techniques, 2, rue de la Houssinière, 44322 Nantes Cedex 03, France.

9 ²Université de La Réunion, 15 avenue René Cassin, 97444 Saint Denis Cedex, La Réunion, France.

10 ³Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India.

11 ⁴INSERM UMR_S 1134, DSIMB team, Laboratory of Excellence, GR-Ex, Université Paris Diderot - Sorbonne Paris
12 Cité, INTS, 6, rue Alexandre Cabanel, 75739 Paris Cedex 15, France.

13 ⁵DSIMB, INSERM, UMR S-1134, Laboratory of Excellence, GR-Ex, Université de La Réunion, Faculty of Sciences
14 and Technology, 97444 Saint Denis Cedex, La Réunion, France.

15 ⁶PEACCEL SAS, 6 square Albin Cachot, 75013 Paris, France.

16 [#]Present address : National Cancer Institute, Bethesda, MA, USA

17

18 ***Corresponding author:**

19 Prof Bernard Offmann

20 Université de Nantes, Unité Fonctionnalité et Ingénierie des Protéines (UFIP), UMR 6286 CNRS, UFR Sciences et
21 Techniques, 2, rue de la Houssinière, 44322 Nantes Cedex 03, France.

22 Phone : +33 2 5112 5721,

23 Email : bernard.offman@univ-nantes.fr

24

25 **Running title:** Prediction of protein backbone conformation

26 **Abstract**

27

28 Libraries of structural prototypes that abstract protein local structures are known as structural
29 alphabets and have proven to be very useful in various aspects of protein structure analyses and
30 predictions. One such library, Protein Blocks (PBs), is composed of 16 standard 5-residues long
31 structural prototypes. This form of analyzing proteins involves drafting its structure as a string of
32 PBs. Thus, predicting the local structure of a protein in terms of protein blocks is a step towards
33 the objective of predicting its 3-D structure. Here a new approach, kPred, is proposed towards this
34 aim that is independent of the evolutionary information available. It involves (i) organizing the
35 structural knowledge in the form of a database of pentapeptide fragments extracted from all
36 protein structures in the PDB and (ii) apply a purely knowledge-based algorithm, not relying on
37 secondary structure predictions or sequence alignment profiles, to scan this database and predict
38 most probable backbone conformations for the protein local structures.

39 Based on the strategy used for scanning the database, the method was able to achieve efficient
40 mean Q_{16} accuracies between 40.8% and 66.3% for a non-redundant subset of the PDB filtered at
41 30% sequence identity cut-off. The impact of these scanning strategies on the prediction was
42 evaluated and is discussed. A scoring function that gives a good estimate of the accuracy of
43 prediction was further developed. This score estimates very well the accuracy of the algorithm (R^2
44 of 0.82). An online version of the tool is provided freely for non-commercial usage at
45 <http://www.bo-protscience.fr/kpred/>.

46

47 **Keywords:** structural alphabet, protein blocks, protein backbone, pentapeptides, database,
48 prediction algorithm

49

50 **Introduction**

51 Knowledge of protein structure considerably helps towards understanding protein function. The
52 Protein Data Bank (PDB) that serves as the central repository of knowledge for the protein
53 structural biology community contains more than 125,000 protein structures and its growth has
54 been considerable in the past decade¹. This number is however still far below the ~70 million
55 protein sequences referenced in UniProt database². Hence it is at stake to find methods to bridge
56 this considerable gap. Computational methods for predicting protein secondary and tertiary
57 structure have persistently tried to fill it. In this paper, we explore the ability of a structural
58 alphabet based prediction method to fulfill in part this role.

59 Since the seminal works by Kabsch and Sander in 1984³, one of the most popular and rewarding
60 computational method to predict and analyze protein structures is by breaking them down to their
61 constituent parts in the so-called fragment-based approach. Multiple fragment libraries have been
62 developed so far and they differ in the number of fragments, the length of the fragments, the
63 methods used for clustering and the criteria used for clustering. The first fragment library was
64 developed by Unger and co-workers⁴. There are reviews that give a good overview of the
65 different fragment libraries developed since then^{5,6}. Also referred to as structural alphabets (SAs),
66 these have shed some light on the sub-secondary structure level intricacies in proteins⁷. By
67 identifying redundant structural fragments found in proteins, structural alphabets help in
68 abstracting protein structures accurately. Such collections of fragments have also been used in
69 methods that attempt to reconstitute protein structures⁸⁻¹⁰.

70 In that respect, a SA called *protein blocks* (PBs) was developed for the purpose of describing and
71 predicting the local backbone structure of proteins^{11,12}. This SA accounts for all local backbone
72 conformations in protein structures available in the Protein Data Bank (PDB). Since then, PBs
73 have been used in various applications¹¹: for structural motif identification¹³⁻¹⁵, structural
74 alignments^{16,17} and fold recognition^{18,19}. There have also been various efforts to use PBs to predict

75 protein local structure. These approaches are based on the Bayes theorem^{9,12} support vector
76 machines²⁰⁻²² and neural networks²³. Some of these methods have used prior predictions of
77 classical three state secondary structures (svmPRAT²² uses YASSPP²⁴, SVM-PB-Pred²¹ uses
78 GOR²⁵ and Etchebest et. al.²⁶ use PSI-PRED²⁷) and sequence alignment profiles like position
79 specific scoring matrices (PSSMs) are used by LOCUSTRA²⁰, SVM-PB-Pred²¹ and Dong and
80 coworkers methodology²³. The currently available web-based tools that can predict local structure
81 in terms of protein blocks are LocPred²⁸ and SVM-PB-Pred²¹. The former implements a Bayesian
82 methodology and the latter is SVM-based.

83 In this work we describe PB-kPRED, a purely fragment and knowledge-based approach to predict
84 local backbone structure of proteins in terms of protein blocks and a web-based tool that
85 implements the method. In essence, it takes no other inputs than the amino acid sequence of a
86 query and interrogates a database of pentapeptides extracted from protein structures, without
87 using evolutionary information. It returns the predicted local structures of the polypeptide chain
88 in the form of a sequence of protein blocks. Very importantly, PB-kPRED also implements a
89 scoring function that efficiently auto-evaluates the quality of the prediction.

90 **Methods**

91 *Dataset*

92 All the protein chains from PDB¹ were segregated into clusters culled at 30% sequence identity
93 using the BLASTClust algorithm²⁸ resulting in a collection of 15,544 clusters. The dataset which
94 we set-up comprises of 15,554 protein chains each corresponding to the best representative
95 structure available from each of these clusters and is hereafter termed as “PDB30 dataset”.
96 Preference was given to crystallographic structures over NMR and electron microscopy structures
97 and also preferring better resolution and lowest R-value structures. Out of these 15,544 structures,
98 14,207 are crystallographic structures, 1,128 are from NMR experiments and 209 are solved by
99 electron microscopy. Further, chains smaller than 100 residues were filtered out. We preferred to
100 keep the NMR and EM structures, as we wanted to investigate if the experimental method impact
101 on the quality of the predictions. For each of these 15,544 proteins the subsets of PDB that were
102 homologous at 30%, 40%, 50%, 70%, 90%, 95% and 100% as reported by BLASTClust were
103 also calculated in order to implement the “*Hybrid method with noise filtering*” scheme.

104

105 *Protein Blocks*

106

107 The set of protein blocks (PBs) is a structural alphabet composed of 16 structural prototypes each
108 representing backbone conformation of a fragment of 5 contiguous residues^{11,12}. The 16 PBs are
109 represented by the letters *a* to *p* and were identified from a collection of 228 non-redundant
110 proteins. Clustering these pentapeptides was based on the 8 dihedral angles (ψ_{i-2} , ϕ_{i-1} , ψ_{i-1} , ϕ_i , ψ_i ,
111 ϕ_{i+1} , ψ_{i+1} , ϕ_{i+2}) that define their local backbone conformation. An unsupervised learning algorithm
112 (Kohonen algorithm) was used to arrive to an unbiased classification of the dihedral vectors and

113 to the definition of standard dihedral angles for each PB. Protein blocks are assigned on the basis
114 of the dissimilarity measure called root mean square deviation on angular values (*rmsda*) between
115 observed dihedral angles and the standard dihedral angles for the 16 PBs. The PB with lowest
116 *rmsda* is assigned to the central residue of the pentapeptide region. The choice of fragment size as
117 5 and library size as 16 for the PBs was because 5 consecutive residues capture well the local
118 contacts in regular secondary structures (α -helices and β -strands) and 16-library size is a good
119 balance between the specificity and sensitivity of predictions¹².

120 All the 15,544 protein chains from PDB30 dataset were encoded into their corresponding protein
121 blocks sequences (PB sequences) after comparing their backbone ϕ and ψ torsion angles with the
122 corresponding standard torsion angles for the 16 PBs¹¹ using an in-house developed Perl script.
123 Sequence of PBs as observed in crystal and NMR structures were later used as a reference to
124 assess the accuracy of predicted PB sequences.

125

126 *Database of pentapeptide conformations from protein structures*

127

128 A database of pentapeptide conformations (PENTAdb) was developed using known 3-D
129 structures of proteins. PENTAdb is essentially the entire structural information contained in the
130 PDB, broken down into chunks of pentapeptides. A sliding window of 5 residues was used to
131 extract structural features for every overlapping pentapeptide of a polypeptide chain. The dihedral
132 vector associated with the five consecutive residues that is required to assign PBs as described in
133 the previous section was obtained from the DSSP²⁹ program. All the information was stored as a
134 MySQL relational database. PENTAdb is maintained up-to-date; the update frequency
135 corresponds to the weekly updates of PDB. The protein chains from which the pentapeptides are
136 extracted are filterable at 30%, 40%, 50%, 70%, 90%, 95% and 100% sequence identity
137 thresholds.

138 *Prediction scheme*

139 The overall scheme for predicting the local structure in terms of PBs is based on querying the
140 PENTAdb database for every constitutive pentapeptides of a query protein sequence using a
141 sliding window of 5 residues (Figure 1a). Hits from the database are reported as predicted protein
142 blocks (PBs). Predicted PBs are assigned to the central residue of each query pentapeptide. The
143 prediction results are presented at different levels of refinement. The prediction in the coarsest
144 form consists of the list of all the possible PBs for a particular pentapeptide of the query protein
145 sequence. This is the case when multiple hits from PENTAdb database are obtained for a
146 particular query pentapeptide (Figure 1b). The multiple hits correspond to the different
147 conformations, which the pentapeptide has been seen to adopt in protein structures (Figure 1b).
148 When the query pentapeptide is not found in PENTAdb, the information available for the
149 tetrapeptides covering the first four residues with a wildcard for the fifth position was used
150 (Figure 1b) to identify the list of possible PBs with first 4 amino acid residues matching this
151 query. The position of wildcard did not influence the outcome of the results (data not shown). The
152 list of hits thus obtained is referred as *all possible PBs*. This list serves as a framework from
153 which the most probable PB sequence is predicted.

154

155 [Fig. 1 about here]

156

157 Two methods were explored to predict the optimal PB sequence within the list of all the possible
158 PBs obtained after querying the database. The first method, termed as *majority rule method*, is
159 purely probabilistic and consists of simply picking up the most frequently observed PB for each
160 query pentapeptide. As shown in Figure 2, it corresponds to the PB that has highest S1 score,
161 where S1 scores are simply the raw counts of all possible PBs reported by PENTAdb database for

162 the query pentapeptide. In cases when there is no decisive majority (two or more equi-probable
163 PB), both of them are reported as predictions.

164

165 However, it is known that the structure adopted by a short peptide can be highly dependent on its
166 local environment³. A second method that integrates contextual information was hence developed
167 and is hereafter termed as *hybrid method*. Here, to predict the local structure of a pentapeptide,
168 the information about the structural status (in terms of PBs) of the two immediately adjacent and
169 overlapping pentapeptides (preceding and succeeding) is also taken into account (see Figure 2). It
170 requires a normalized frequency look-up table for observed motifs of 3 consecutive PBs also
171 termed as tri-PBs (see “additional methods” section of the supplementary material). For each
172 query pentapeptide, in complement to the calculated S1 score, an additional S2 score is calculated
173 as follows. A list of all possible combinations of three successive PBs (tri-PB motifs) is built. This
174 is derived from the list of *all possible PBs* for the query pentapeptide and for its two adjacent
175 pentapeptides (Figure 2). For each possible tri-PB motif, their normalized frequencies (“odds” in
176 Figure 2) are looked up in the tri-PB normalized frequency table. S2 scores are calculated through
177 the summation of the odds of tri-PB motifs that have a common PB in the central position (Figure
178 2). The predicted PB for the query pentapeptide is determined after multiplying S1 scores by their
179 corresponding S2 scores and taking the highest value among these products (Figure 2). This
180 approach is called the *hybrid method* because it combines the *majority rule method* with
181 contextual information in the prediction process.

182

183 [Figure 2 about here]

184 *Evaluating PB-kPRED using different subsets of PENTAdb*

185 Two evaluation schemes were developed to benchmark the PB-kPRED methodology. As

186 mentioned above, the query dataset used here constituted of the 15,544 proteins from the PDB30
187 dataset. The schemes relied on the ability to control which subsection of PENTAdb will be
188 accessible to the prediction algorithm for every query. For example, allowing only pentapeptides
189 in PENTAdb from non-homologues to be accessible by the prediction algorithm emulates a
190 scenario of attempting to predict the local structure of a protein with no homologue of known
191 structure used. On the other hand, as in the case of other local structure prediction
192 methods^{12,20–22,26}, it can be advantageous to have the ability to privilege information from
193 homologous structures when these are available to predict the local structure of a query protein.
194 Such a scheme can be emulated by allowing only pentapeptides in PENTAdb from closest
195 detectable homologues to be accessible by the search algorithm.

196 In first instance, the prediction methodology was assessed with increasing sequence identity cut-
197 offs ranging from 30%, 40%, 50%, 70%, 90%, 95% to 100%, named experiments A1-A8 (see
198 Figure 3a). This scheme is subsequently termed as “*without noise filtering scheme*”.

199 In second instance, an alternative assessment scheme hereby called the “*with noise filtering*
200 *scheme*” was applied to further assess the PB-kPRED methodology (experiments B1-B8, see
201 Figure 3b). It aimed at evaluating how privileging information from close homologues, when
202 available, contributed to the quality of the predictions. In brief, the algorithm initially searches for
203 a pentapeptide among the closest homologues first. If the search finds a hit, then the hit is used
204 for the prediction; otherwise the search space is increased to include the immediately next level of
205 more distantly related homologues. This process is repeated until a hit is obtained. Due to this
206 process of introducing more distant homologues in a conditional fashion, wrong pentapeptides
207 (noise) from PENTAdb were potentially filtered out, hence the name *with noise filtering scheme*.

208 In all the cases, care was taken to exclude the pentapeptides from the query proteins themselves.

209

210 [Figure 3 about here]

211

212 Reducing the PB predictions into a binary outcome permits the use of classical Mathews
213 correlation coefficient (MCC) to compare our predictions to a random choice. MCCs for the 16
214 PBs were evaluated based on a confusion matrix similar. For each PB, MCC was calculated
215 according to Equation 1.

216

$$217 \quad MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (1)$$

218

219 *A scoring function to estimate the accuracy of the predictions*

220 A probabilistic scoring function was developed for the *a posteriori* analysis of the predicted PB
221 sequence through namely the analysis of its content in penta-PB motifs, with the objective of
222 providing a measure of how accurate PB-kPRED was performing. The principle of the analysis
223 relies on the fact that not all penta-PBs are commissioned by proteins at the same frequency.
224 Indeed, many successions of 5 consecutive PBs are highly improbable because they are
225 geometrically not allowed as explicated by the Ramachandran rules. The probabilistic function is
226 hence based on the look-up table of normalized frequencies of successive penta-PB motifs
227 observed in a non-redundant set of protein structures (see “additional methods” in supplementary
228 material). In brief, using a sliding window of 5 consecutive PBs (penta-PB motif) along the
229 predicted PB sequence, the normalized frequencies of all penta-PB motifs were looked-up in the
230 penta-PB frequency table. The logarithm of these normalized frequencies were then summed and
231 divided by the length of the predicted PB sequence to generate an *accuracy score* (A) as shown
232 here:

233

$$A = \frac{\sum_{i=1}^{l-4} \log(N_i)}{l} \quad (2)$$

234 where A is the *accuracy score* for a predicted PB sequence, l is the length of the PB sequence, N
235 is the normalized frequency of the penta-PB motif observed at window position i in the PB
236 sequence. Since an overlapping sliding window of five consecutive PBs is used, the total number
237 of penta-PB motifs (*i.e* the number of windows) is $l-4$. In the case a particular penta-PB motif has
238 a null value in the frequency table (*i.e* it is never observed), a penalty of -5 was instead added to
239 the score.

240 **Results**

241

242 *PENTAdb, a database of pentapeptides from protein structures*

243 A total of 68.84 million pentapeptides obtained from the 0.26 million protein chains and their
244 corresponding local structure represented as one of the 16 PBs were obtained and stored in
245 PENTAdb. Of these 68.84 million pentapeptides, 2.26 million are unique which represents 70.9%
246 of the total number theoretically possible 3.2 million (20^5) pentapeptides. The content of the
247 database accessible to PB-kPRED at these threshold values is given in Table 1. There is a 32-fold
248 decrease (from 68.62 to 5.13 million) in the number of pentapeptides in PENTAdb when PDB
249 chains not sharing more than 30% sequence identity are considered. Nevertheless, there is only a
250 1.3 fold decrease in the number of unique pentapeptides present in PENTAdb at this threshold.

251

252 [Table 1 about here]

253

254 *Not all possible tri-PB combinations are observed in known protein structures*

255 Out of all the theoretically possible 4,096 (16^3) tri-PBs, a total of 1,375 (i.e 33.5%) were never
256 observed in the non-redundant PDB30 dataset. Likewise, out of all the 1.04 million (16^5)
257 theoretically possible penta-PB motifs, only 40,130 (3.8%) were observed in the PDB30 dataset.
258 These results are indicative of the possibility that many combinations of three or five consecutive
259 PBs are stereochemically unfavorable. The distributions of penta-PB motifs at other sequence
260 identity cut-offs i.e. 40%, 50%, 70%, 90%, 95%, 100% and the entire PDB were also computed
261 (Supplementary Table 1). Towards higher sequence identity cut-offs, there was a steady increase
262 in the penta-PB coverage. But this comes at the price of the addition of redundant data. Still the
263 entire PDB covered less than 10% of the total penta-PB space. For calculating the *accuracy score*,
264 the penta-PB frequency table derived from the PDB30 dataset was used even if it contained only

265 40,130 penta-PB motifs. This might seem to be a small fraction but this was sufficient to
266 efficiently score PB sequences (see below).

267

268 *Completeness of PENTAdb for knowledge-based prediction*

269 A quantitative assessment of how often the correct PB can be found in the list of all possible PBs
270 reported for every query protein was performed. This represents the theoretical highest prediction
271 rate attainable for a query protein using the proposed knowledge-based approach. To this end, for
272 every query protein sequence, different portions of the pentapeptide database were made
273 accessible to the prediction algorithm. This is manageable, thanks to the hierarchical clustering at
274 different sequence identity levels by the BLASTClust algorithm²⁹. For each of the 15,544
275 unrelated query protein sequences (PDB30 dataset), only pentapeptides coming from a subset of
276 the PDB that shared sequence identities below an indicated cut-off values (from 30% to 100%)
277 and excluding the query itself were made accessible to PB-kPRED for prediction (see Table 1 for
278 size of the database for each subset). The results are detailed in Table 2. At 30% sequence identity
279 cut-off, the correct PB was found in 71.4% of the case and the success rate increased to 77.3%
280 when only “homologues” sharing 100% sequence identity to the queries were filtered out. When
281 full PDB was used (but excluding the query) as a database, the percentage times the correct PB is
282 in the list of all possible PBs topped to 99.93%. The PB-wise breakdown of these values are
283 further detailed in supplementary Table 2.

284

285 [Table 2 about here]

286

287 *Prediction accuracies*

288 The average prediction accuracies for the PDB30 query proteins using the *majority rule method*
289 and the *hybrid method* using the classic scheme for querying the database are given in Table 3.

290 When homologues sharing $\geq 30\%$ sequence identity with each of the queries were removed from
291 the database, PB-kPRED performed with an average Q16 accuracy of 39.2% and 40.8% for the
292 *majority rule method* and *hybrid method* respectively (Table 3). Surprisingly, the effect of
293 enlarging the database to include closer homologues sharing $< 95\%$ sequence identity with the
294 queries improved only marginally the prediction accuracies reaching on average 40.4% and
295 42.4% for *majority rule method* and *hybrid method* respectively. Accuracy topped to 58.0% and
296 54.6% respectively when full PDB (excluding the query itself) was used as database for
297 prediction. This overall gain in accuracy is due to an incremental increase of accuracy across all
298 the 16 PBs.

299

300 [Table 3 about here]

301

302 As an attempt to improve the prediction rates, the hybrid method was tested using the *noise*
303 *filtering scheme* for querying the database whereby, for each query pentapeptide, data in
304 PENTAdb only coming from closest homologues was queried first (see Figure 3). Results are
305 detailed in Table 4. When compared to the *without noise filtering scheme* (Table 3), the prediction
306 rates improved to reach a maximum of 66.3%. Interestingly, for experiments B2 to B7 where
307 closest homologues to be queried first are in the range of $< 40\%$ to $< 95\%$ sequence identities, the
308 predictions remained high at a level of about 61.6%. Only in experiment B8 the prediction
309 accuracy rate dropped to 40.8%. This experiment is in fact identical to the one featured for $< 30\%$
310 threshold shown in Table 3 using the hybrid method and the *without noise filtering scheme* for
311 querying the database.

312

313 [Table 4 about here]

314

315 All results further detailed hereafter are concerned with data obtained in experiment B1 where
316 hybrid method was applied using the *noise filtering scheme* for querying PENTAdb and where
317 best predictions were obtained.

318

319 The distribution of the prediction accuracies for experiment B1 (see Table 4) shows a bimodal
320 distribution (see Figure 4). A spike in frequency is observed at the >80% range representing the
321 set of queries which have closely related proteins of known structure available in the PDB and for
322 which the method is able to perform extremely well. At the other end of the spectrum, there is an
323 almost normal distribution with an average around the 35%-40% accuracy range. Hence, the
324 mean falls in between these two at 66.31% accuracy. This distribution did not substantially vary
325 when homologues sharing less than 100% to 40% sequence identity to the query corresponding to
326 experiments B2 to B7 respectively (see Figure 3) were queried first (data not shown). However,
327 once the twilight zone of 30% sequence identity is crossed, the accuracy distribution drastically
328 changes to that of a unimodal distribution with a very sharp peak at the 40% range and gradually
329 tapering tail towards the higher accuracies (supplementary Figure 1).

330

331 [Fig 4 about here]

332

333 The accuracy by the *hybrid method* using the *noise filtering scheme* was compared to the *majority*
334 *rule method* (Figure 5). As shown by the data points below the diagonal, the *hybrid method*
335 performed significantly better than the *majority rule method* for a total of 8,195 cases (52.7%) out
336 of the 15,544 protein queries. For remaining 7,245 cases, the *majority rule method* performed
337 slightly better than the *hybrid method*.

338

339 [Fig 5 about here]

340

341 *PB predictions*

342 Results from the best performing condition (experiment B1 featured in Table 4 and Figure 3b)
343 were further analyzed for the PB-wise prediction rates and compared with published rates from
344 other methods (Table 5). The rates are heterogeneous across the 16 PBs. Top two best-predicted
345 PBs by PB-kPRED were PB *m* and PB *a*, with accuracies of 75.9% and 67.2%, respectively. On
346 the other hand, the two most badly predicted PBs by PB-kPRED were PB *j* and PB *g* with
347 prediction rates of 49.9% and 43.5% respectively. Analysis of the corresponding confusion matrix
348 (see supplementary Table 3) shows that, the prediction algorithm frequently gets confused
349 between the PBs *c* and *d*. PB *c* is wrongly predicted as PB *d* almost 31,000 times (22.4%). The
350 vice-versa, PB *d* being predicted as PB *c* is more than 42,000 times (16.9%). These PBs are in
351 fact highly related (i) as seen from a pure structural point of view (low *rmsda* and similar
352 transitions)²⁷ and (ii) as they have been seen to be highly interchangeable thanks to PB
353 substitution matrix^{31,32}. As some PBs are highly similar, it is possible to relax the assessment, *i.e.*
354 considering two PB series as equivalent. With such relaxed criteria, the accuracy increases from
355 66.31% to 68.87% (a 2.56% gain on average). Interestingly, significant increases in accuracies
356 were observed for PB *g* (from 43.5% to 67.4%) and for PB *j* (from 49.98% to 67.2%).

357

358 [Table 5 about here]

359

360 PB-kPRED globally outperformed two other PB prediction methods (see Table 5). Its predictions
361 were better for all the 16 PBs when compared to the Bayes method and better than almost all PBs
362 when compared to LOCUSTRA. Only PBs *d* and *m* were better predicted by this latter method²⁰.

363

364 A MCC close to +1 indicates a good agreement between the observed and the predicted outcomes

365 and a MCC of close to -1 otherwise. For our analysis all the 16 PBs had MCCs between 0.5 and
366 0.7. PBs *a* and *m* were close to 0.7, PB *g* at 0.51 and the remaining fluctuated around the 0.6
367 mark. The sensitivity and specificity ranges were 0.4-0.7 and 0.9-1.0 respectively. A common
368 pattern is observed in the case of PBs corresponding to the regular secondary structure elements
369 (PBs *d* and *m*): in both these cases, the sensitivity values peak while the specificity values
370 plummet. Although the sensitivity values varied between 0.4 and 0.8, the specificity values were
371 consistently above 0.9 indicating that the method was able to achieve a very high true negative
372 rate.

373

374 *Measure of accuracy*

375 A probabilistic scoring function was developed for the *a posteriori* analysis of the predicted PB
376 sequences so as to provide a measure of how accurate the *hybrid method* using the *noise filtering*
377 *scheme* was performing. An assessment of the scoring function is provided in Figure 6. It shows
378 that the score is correlated with the accuracy of the prediction with a Pearson's correlation
379 coefficient of 0.82 (Figure 6a). The two distinct clusters of data points correspond to those
380 featured in the histogram in Figure 4. As a further assessment of the scoring function, the scores
381 for the predicted PB sequences were compared with the scores for the actual PB sequences
382 (Figure 6b). It shows that in case of more accurate predictions (rates above 60%), the two scores
383 correlated very well (red points along the diagonal in Figure 6b) with both score values mostly
384 ranging between +1 and +3. In the case of less accurate predictions (rates below 60%), the two
385 scores were no more correlated (green dots below the diagonal in Figure 6b) and scores for
386 predicted PB sequences ranged mostly between -2 and +1.

387

388 [Fig. 6 about here]

389

390 *Case studies*

391 Here were considered the predictions for 5 specific cases to look at the strengths and limitations
392 of the PB-kPRED algorithm namely in presence or absence of homologues of known structure.
393 Results are reported in Table 6 and further described below. These case studies correspond to
394 counter intuitive prediction instances where (i) prediction accuracy is high despite not having any
395 close homologues and (ii) prediction accuracy is low despite having sequences of PBs of closely-
396 related proteins in PENTAdb.

397

398 [Table 6 about here]

399

400 Regarding prediction in employing information from homologues of known structure, three
401 contrasting cases were studied. The first case relates to chain A of a hypothetical DNA binding
402 protein from *Salmonella cholera* (PDB id 2HX0_A) which has a homologue from *Salmonella*
403 *typhimurium* (PDB id 2NMU) that is 100% identical, 100% accuracy was achieved as shown by
404 the high accuracy score of 2.81. Both structures aligned very well with a RMSD of 0.14 Å
405 (Supplementary Figure 2a). The second case relates to an energy-coupling factor transporter
406 transmembrane protein EcfT from *Lactobacillus brevis* (PDB id 4HUQ_T) which has two other
407 “homologues” (PDB id 4RFS_T and 4HZU_T) that are 100% identical to the query. Here, the
408 prediction accuracy is 73.4% only with an accuracy score of 1.31. 3-D structural alignment with
409 these two “homologues” resulted in RMSDs 1.41 Å and 1.90 Å respectively displaying some
410 structural variations (Supplementary Figure 2b) despite being 100% identical at the amino acid
411 sequence level. These structural variations were due to rigid body movement. The third case is
412 chain A of a pyrimidine deaminase / uracil reductase from *Thermotoga maritima* (PDB id
413 2HXV_A) which had only ten very distantly related proteins that shared less than 30% sequence

414 identity in the PDB. Prediction rate is even lower here with accuracy reaching a value of 39.1%
415 as shown by the low accuracy score of 0.30.

416

417 As for predictions in absence of homologues of known structure, two contrasting cases were
418 studied. The first case is about a human hydroxysteroid dehydrogenase and the second case is a
419 membrane protein associated with Ecf transporter from *Lactobacillus brevis*. The prediction
420 performed quite well in the first case with an accuracy of 75.4% as shown by the high accuracy
421 score of 1.99 while in the second case, the prediction almost completely failed with the accuracy
422 of only 9.37% and also shown by the unfavorable accuracy score of -0.28.

423 *Implementation of the PB-kPRED methods as a web-tool*

424 The PB-kPRED methodology has been implemented as a web-tool that is freely available to the
425 community at <http://www.bo-protscience.fr/kpred/>. Both *majority rule* and *hybrid methods*
426 without the *noise filtering scheme* for querying the database have been implemented. The tool
427 provides a predicted PB sequence for each query amino acid sequence and also provides the
428 *accuracy score* that serves as an *a posteriori* estimation of the prediction accuracy. In case the
429 prediction score value is below -1, the prediction accuracy cannot be estimated and the user is
430 notified. Users can provide multiple query protein sequences. All results are downloadable as
431 FASTA formatted flat files. Optionally when submitting numerous query sequences, the user can
432 provide an email address to which a notification will be sent when the job is completed.

433 Discussion

434 The exponential growth in the structural knowledge of proteins has warranted the necessity of
435 competent knowledge-based prediction algorithms for local structure prediction. At the level of
436 short protein segments like pentapeptides, this increase in structural knowledge invariably brings
437 with it an unprecedented signal to noise ratio for deciding on the most probable local
438 conformations. Indeed, it is well established that similar pentapeptides can adopt different local
439 conformations^{3,33}. This is verified when the content of PENTAdb is inspected.

440

441 Hazout's team along with defining the protein blocks also predicted the local structure in terms of
442 PBs using a Bayesian approach¹². They achieved an accuracy of 34.4% using a 15-residue
443 window and this increased to 40.7% upon supplementing the Bayesian predictor with sequence
444 profiles in the form of *sequence families*. In 2005, Etchebest and colleagues²⁶ used a combination
445 of statistical optimization procedure and improved sequence family data to bump up the accuracy
446 to 48.7%. Incorporating secondary structure predictions from PSI-PRED into the PB prediction
447 process did not contribute much to improve the accuracy *i.e* only 1% gain resulting in 49.9%.
448 Machine learning techniques have also been used to predict protein local structure in terms of
449 PBs. Support vector machine based methods like LOCUSTRA²⁰, svmPRAT²² and SVM-PB-Pred²¹
450 achieve mean accuracies of 61.0%, 67.0% and 53.0% respectively. A dual layer neural network
451 based prediction method achieved 58.5% accuracy²⁴. The most refined version of the PB-kPRED
452 method proposed here, *i.e hybrid method with noise filtering scheme*, outperformed most of the
453 previously developed methods for PB prediction except for svmPRAT where it performed
454 equivalently. Although all the methods evaluated their accuracies on non-redundant sets of
455 proteins, an even comparison is hindered by difference in datasets, varying training regimes for
456 the machine learning methods and different levels of sequence identity used as input in the
457 prediction process. This motivated us to perform a battery of tests on the algorithm to estimate the

458 prediction accuracy when incremental levels of sequence identities are made available in
459 PENTAdb for the prediction (see Table 4). Importantly, to our knowledge, this is the first report of
460 a querying scheme that dynamically filters out, on a per query basis, homologues at different cut-
461 off values so that the portion of the PENTAdb that is made accessible for prediction is calculated
462 on the fly. For the each of the 15,544 query sequences of PDB30, 16 experiments were performed
463 amounting a total of 248,704 datasets building. This is computationally intensive and was
464 performed using extensive MySQL querying. Thanks to the *noise filtering* strategy, PB-kPRED
465 was able to efficiently weed out the noise present in the database due to redundancy and hence to
466 narrow down the search in the database to find the most appropriate local structure for a given
467 pentapeptide. Hence, filtering out from the database the pentapeptides from proteins that shared
468 less than 30% sequence identity with the query indeed improves the prediction efficiency.

469
470 When the *majority method* and the *hybrid method* (Figure 5) were compared, two distinct clusters
471 were noticed. Upon further investigating the reason for this distinct clustering, we note that,
472 irrespective of the sequence identity cut-off, the points below the diagonal were found in more
473 populated clusters while the points above the diagonal were found in least populated clusters
474 Hence the *hybrid method* using the *noise filtering scheme* will perform better when there are some
475 closely-related protein structures to look-up to in PENTAdb. In a real-life scenario, this will not
476 be always the case. Indeed, proteins for which we want to predict the structure and which do not
477 have any homologues even at 30% sequence identity are not so uncommon. This brings us to the
478 conclusion that even though overall the *hybrid method* performs better, we cannot ignore the
479 *majority rule method* all together.

480

481 Nonetheless, this method still has room for improvement as it can be seen from the values in
482 Table 2. The list of all possible PBs reported by the PB-kPRED algorithm after querying

483 PENTAdb database indeed shows that the good PB was present in more than 70% of the cases.
484 However, owing to the scoring functions S1 and S2 (Figure 2), the decision rules implemented in
485 both *majority rule* and *hybrid methods* failed to pick up these good PBs as predictions in several
486 instances.

487

488 Interestingly, once the local backbone of a protein was predicted in the form of a PB sequence, we
489 were able to provide an *a posteriori* assessment of how accurate was the prediction. The method
490 used here to achieve this relied on the simple idea that successions of PBs should follow the rule
491 that not all combinations of PBs would be allowed. This intuition turned out to be correct since
492 there was a remarkable correlation between the score and the accuracy of the predictions.
493 Noteworthy, the *accuracy scores* for actual (native) PB sequences are overwhelmingly distributed
494 between +1 and +3, while poorly predicted PB sequences have scores below +1. This scoring of
495 PB sequences could also serve as an indicator towards improving predictions. Because the
496 calculation of the score of a PB sequence is very fast, one could imagine implementing a score-
497 guided optimization procedure to climb the prediction accuracy gradient using Monte-Carlo or
498 genetic algorithms for example.

499

500 The case studies documented in this work (see Table 6) indicate that the relationship between
501 local structure predictability and the number of homologues of the query available in PDB are not
502 very straightforward. Optimistically, in spite of not having any homologues, the PB-kPRED
503 algorithm can perform a good prediction if the pentapeptides constituting the query adopt
504 consensus local structures for the respective pentapeptides. Two such examples were provided but
505 with contrasting outcomes, one achieving good accuracies and the other failing to predict
506 correctly the PB sequence. Interestingly, the *accuracy scores* provided by our scoring function
507 helped to reliably differentiate one prediction from the other. On the other hand, even if a query

508 has multiple homologues in the PDB, its prediction accuracy will take a hit if the homologues are
509 contrasting structural analogues of the query. For example, the activation of human pancreatic
510 lipase involves considerable conformational transition in the form of a 'lid movement'. The
511 hypothetical prediction case when the query is the 'lid open form' and PENTAdb has
512 pentapeptides from the 'lid closed form' would confuse the prediction algorithm despite both the
513 forms of lipase being identical in amino acid sequences. Hence these case studies establish two
514 take home messages: (i) there are exceptions to the general observation that the presence of
515 homologues improves the prediction accuracy of PB-kPRED and (ii) the *accuracy score* used to
516 evaluate the predictions is a reliable gauge for estimating the accuracy of the method as illustrated
517 in Figure 6.

518

519 PB-kPRED web-server could form a vital link in the pipeline of PB based structure analysis tools.
520 Namely, it can be bridged with PB-based fast structure comparison tools like iPBA¹⁷ and
521 PBAalign¹⁶ and help to mine for similar structures and map the fold space. It can also be used to
522 predict the occurrence of structural motifs in protein sequences. Indeed, the alpha version of the
523 server which was made available on-line earlier, has already been used by some research groups
524 for the structural characterization of RNA binding sites in protein structures and predicting
525 proteins sequences that contain RNA binding sites^{34,35} and also in predicting β -turns and their
526 types³⁶.

527

528 The web-based tool currently does not feature the *hybrid method with noise filtering scheme*
529 because it would require running an instance of BLASTClust on every query. We plan to
530 implement this functionality in a future improvement to the tool.

531

532 **Supplementary material**

533 Additional methods file: additional_methods.pdf

534 Supplementary tables file: supplementary_tables.pdf

535 Supplementary figures file: supplementary_figures.pdf

536

537 **Acknowledgements**

538 Authors thank Tristan Riailand and Sara Bachiri for providing technical support in the
539 development of the PB-kPRED web server. This work was supported by the Région Réunion and
540 the Fond Social Européen [grant no. 20131528] to IV. This work was in part supported by Conseil
541 Régional des Pays de la Loire in the framework of GRIOTE project. AdB and FC acknowledge
542 grants from the Ministry of Research (France), National Institute for Blood Transfusion (INTS,
543 France), National Institute for Health and Medical Research (INSERM, France) and labex GR-
544 Ex. The labex GR-Ex, reference ANR-11-LABX-0051 is funded by the program “Investissements
545 d’avenir” of the French National Research Agency, reference ANR-11-IDEX-0005-02. AdB
546 acknowledge supports by University Paris Diderot, Sorbonne, Paris Cité (France), FC
547 acknowledge supports by Université de La Réunion, Faculty of Sciences and Technology. NS and
548 AdB acknowledge to Indo-French Centre for the Promotion of Advanced Research / CEFIPRA
549 for collaborative grant (number 5302-2). Research in NS laboratory is also supported by
550 Department of Biotechnology, Government of India. NS is a J.C. Bose National Fellow.

551

552 *Conflict of Interest:* BO and FC are co-founders of PEACCEL SAS.

553

554 **References**

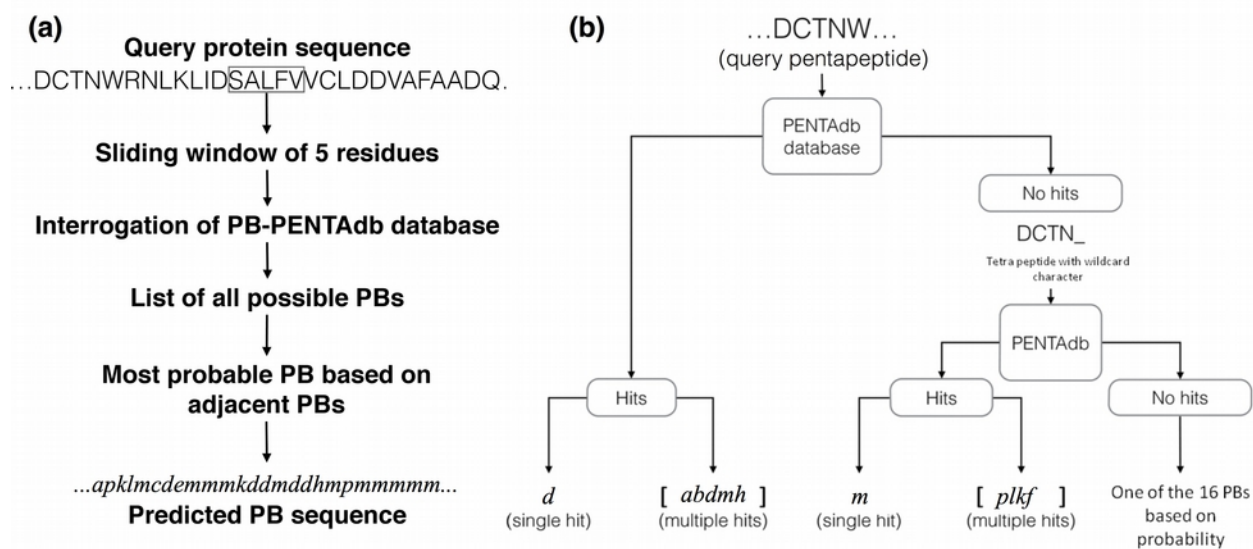
- 555 1. Rose PW, Prlić A, Bi C, Bluhm WF, Christie CH, Dutta S, Green RK, Goodsell DS,
556 Westbrook JD, Woo J, Young J, Zardecki C, Berman HM, Bourne PE, Burley SK
557 (2015) The RCSB Protein Data Bank: views of structural biology for basic and applied
558 research and education. *Nucleic Acids Res* 43:D345–D356.
- 559 2. Uniprot Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res*
560 43:D204–D212.
- 561 3. Kabsch W, Sander C (1984) On the use of sequence homologies to predict protein
562 structure: identical pentapeptides can have completely different conformations. *Proc*
563 *Natl Acad Sci USA* 81:1075–1078.
- 564 4. Unger R, Harel D, Wherland S, Sussman JL (1989) A 3D building blocks approach to
565 analyzing and predicting structure of proteins. *Proteins* 5:355–373.
- 566 5. Karchin R, Cline M, Karplus K (2004) Evaluation of local structure alphabets based on
567 residue burial. *Proteins* 55:508–518.
- 568 6. Offmann B, Tyagi M, de Brevern AG (2007) Local Protein Structures. *Current*
569 *Bioinformatics* 33:165–202.
- 570 7. Tyagi M, Bornot A, Offmann B, de Brevern AG (2009) Protein short loop prediction in
571 terms of a structural alphabet. *Comput Biol Chem* 33:329–333.
- 572 8. Bystroff C, Simons KT, Han KF, Baker D (1996) Local sequence-structure correlations
573 in proteins. *Curr Opin Biotechnol* 7:417–421.
- 574 9. Simons KT, Kooperberg C, Huang E, Baker D (1997) Assembly of protein tertiary
575 structures from fragments with similar local sequences using simulated annealing and
576 Bayesian scoring functions. *J Mol Biol* 268:209–225.

- 577 10. Kolodny R, Koehl P, Guibas L, Levitt M (2002) Small libraries of protein fragments
578 model native protein structures accurately. *J Mol Biol* 323:297–307.
- 579 11. Joseph AP, Agarwal G, Mahajan S, Gelly J-C, Swapna LS, Offmann B, Cadet F, Bornot
580 A, Tyagi M, Valadié H, Schneider B, Etchebest C, Srinivasan N, de Brevern AG (2010)
581 A short survey on protein blocks. *Biophys Rev* 2:137–147.
- 582 12. de Brevern AG, Etchebest C, Hazout S (2000) Bayesian Probabilistic Approach for
583 Predicting Backbone. *Proteins* 287:271–287.
- 584 13. Dudev M, Lim C (2007) Discovering structural motifs using a structural alphabet:
585 application to magnesium-binding sites. *BMC Bioinformatics* 8:106.
- 586 14. Wu CY, Chen YC, Lim C (2010) A structural-alphabet-based strategy for finding
587 structural motifs across protein families. *Nucleic Acids Res* 38:e150.
- 588 15. Schneider B, Cerný J, Svozil D, Cech P, Gelly J-C, de Brevern AG (2014)
589 Bioinformatic analysis of the protein/DNA interface. *Nucleic Acids Res* 42:3381–3394.
- 590 16. Tyagi M, de Brevern AG, Srinivasan N, Offmann B (2008) Protein structure mining
591 using a structural alphabet. *Proteins* 71:920–937.
- 592 17. Joseph AP, Srinivasan N, de Brevern AG (2011) Improvement of protein structure
593 comparison using a structural alphabet. *Biochimie* 93:1434–1445.
- 594 18. Mahajan S, de Brevern AG, Sanejouand Y-H, Srinivasan N, Offmann B (2015) Use of a
595 structural alphabet to find compatible folds for amino acid sequences. *Protein Sci*
596 24:145–153.
- 597 19. Ghouzam Y, Postic G, de Brevern AG, Gelly J-C (2015) Improving protein fold
598 recognition with hybrid profiles combining sequence and structure evolution.
599 *Bioinformatics* 31:3782-9.

- 600 20. Zimmermann O, Ulrich HEH. (2008) LOCUSTRA: Accurate Prediction of Local
601 Protein Structure Using a Two-Layer Support Vector Machine Approach. *J Chem Inf*
602 *Model* 48:1903–1908.
- 603 21. Suresh V, Parthasarathy S (2014) SVM-PB-Pred: SVM based protein block prediction
604 method using sequence profiles and secondary structures. *Protein Pept Lett* 21(8):736–
605 742.
- 606 22. Rangwala H, Kauffman C, Karypis G (2009) svmPRAT: SVM-based protein residue
607 annotation toolkit. *BMC Bioinformatics* 10:439.
- 608 23. Dong Q, Wang X, Lin L, Wang Y (2008) Analysis and prediction of protein local
609 structure based on structure alphabets. *Proteins* 72:163–172.
- 610 24. Karypis G (2006) YASSPP: Better kernels and coding schemes lead to improvements
611 in protein secondary structure prediction. *Proteins* 64(3):575–586.
- 612 25. Garnier J, Gibrat JF, Robson B (1996) GOR secondary structure prediction method
613 version IV. *Methods Enzym* 266:540–553.
- 614 26. Etchebest C, Benros C, Hazout S, de Brevern AG (2005) A structural alphabet for local
615 protein structures: Improved prediction methods. *Proteins* 59:810–827.
- 616 27. Jones DT (1999) Protein secondary structure prediction based on position-specific
617 scoring matrices. *J Mol Biol* 292:195–202.
- 618 28. de Brevern AG de, Benros C, Gautier R, Valadié H, Hazout S, Etchebest C (2004)
619 Local backbone structure prediction of proteins. *In Silico Biol* 4:381–386.
- 620 29. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment
621 search tool. *J Mol Biol* 215:403–410.

- 622 30. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: Pattern
623 recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22(12):2577–
624 2637.
- 625 31. Tyagi M, Gowri VS, Srinivasan N, de Brevern AG De, Offmann B (2006) A
626 Substitution Matrix for Structural Alphabet Based on Structural Alignment of
627 Homologous Proteins and its Applications. *Proteins* 39:32–39.
- 628 32. Joseph AP, Srinivasan N, de Brevern AG (2011) Improvement of protein structure
629 comparison using a structural alphabet. *Biochimie* 93:1434–1445.
- 630 33. Argos P. (1987) Analysis of sequence-similar pentapeptides in unrelated protein tertiary
631 structures. *Strategies for protein folding and a guide for site-directed mutagenesis. J*
632 *Mol Biol.* 197:331-48.
- 633 34. Suresh V, Liu L, Adjero D, Zhou X (2015) RPI-Pred: predicting ncRNA-protein
634 interaction using sequence and structural information. *Nucleic Acids Res* 43:1370–
635 1379.
- 636 35. Luo J, Liu L, Venkateswaran S, Song Q, Zhou X (2017) RPI-Bind: a structure-based
637 method for accurate identification of RNA-protein binding sites. *Sci Rep* 7:614.
- 638 36. Nguyen L, Dang X, Le T, Saethang T, Tran V, Ngo D, Gavrilov S, Nguyen N, Kubo M,
639 Yamada Y and Satou K (2014) Predicting Beta-Turns and Beta-Turn Types Using a
640 Novel Over-Sampling Approach. *Journal of Biomedical Science and Engineering*
641 7:927-940.

642 Figures legends

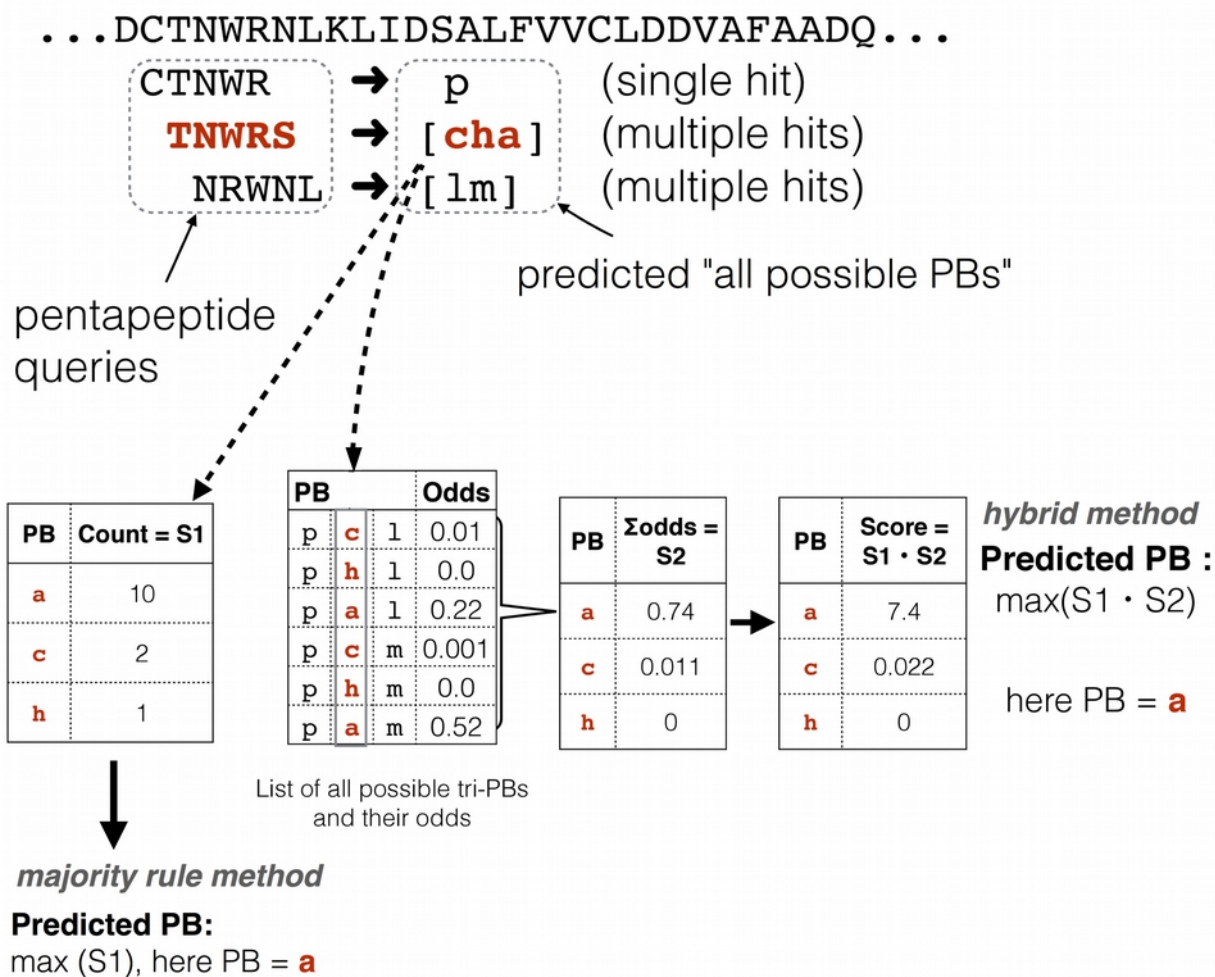


643

644 **Figure 1. The knowledge-based methodology behind PB-kPRED.** (a) Overview of the scheme
645 followed by PB-kPRED for the prediction process. (b) The different outcomes possible when
646 PENTAdb database is queried for a pentapeptide sequence: hits are reported as a single PB or
647 multiple PBs.

648

649



650

651 **Figure 2. Details of the scoring schemes underlying the *majority rule method* and the *hybrid***

652 ***method*.** S1 scores are simply the raw counts of all possible PBs reported by PENTAdb database

653 for a given query pentapeptide. S2 scores are calculated through the summation of the odds of tri-

654 PBs that have a common PB in the central position. For the majority rule method, predictions are

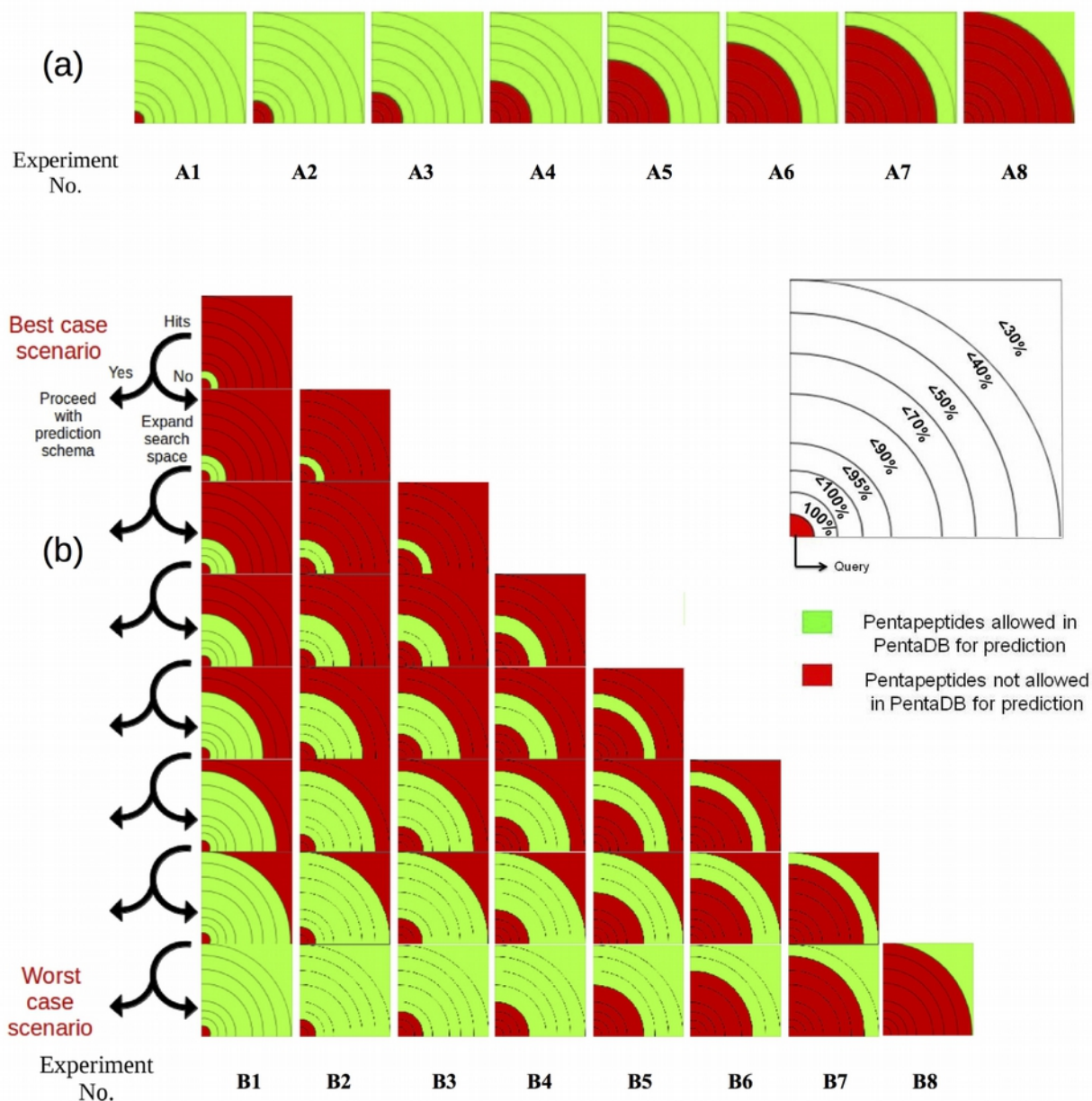
655 based only on the ranking of the S1 scores. For the *hybrid method*, predictions are based on the

656 ranking of the product of scores S1 and S2.

657

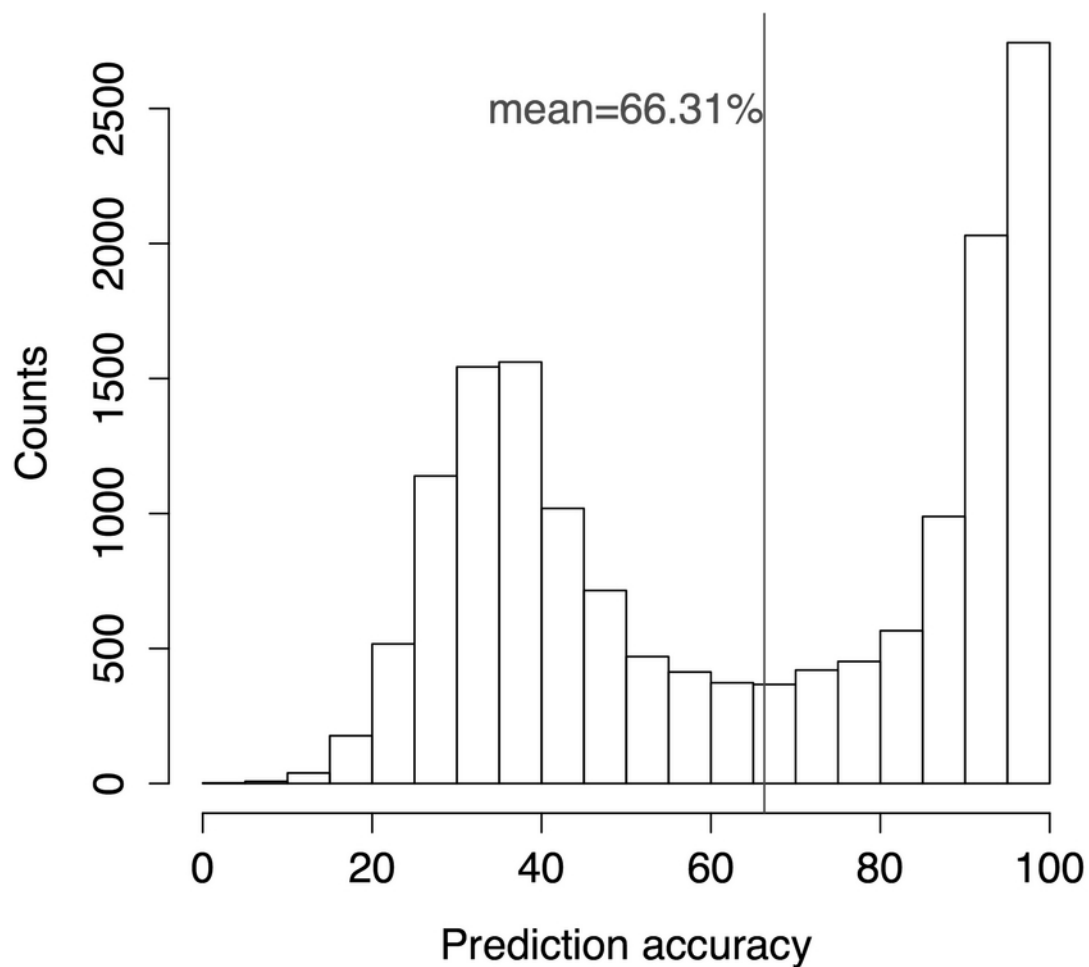
658

659



660

661 **Figure 3. Diagrammatic representations of the schemes used by PB-kPRED for querying**
 662 **PENTAdb with (a) representing the so-called “classic” or “without noise filtering scheme”**
 663 **and (b) representing the “with noise filtering scheme”.** Sections of the database accessible are
 664 indicated in green and those not accessible in red. The sections are delimited by sequence identity
 665 thresholds. In both schemes, eight different experiments (A1 to A8 and B1 to B8) were
 666 performed. See “additional methods file” in supplementary material for detailed legend of this
 667 figure.



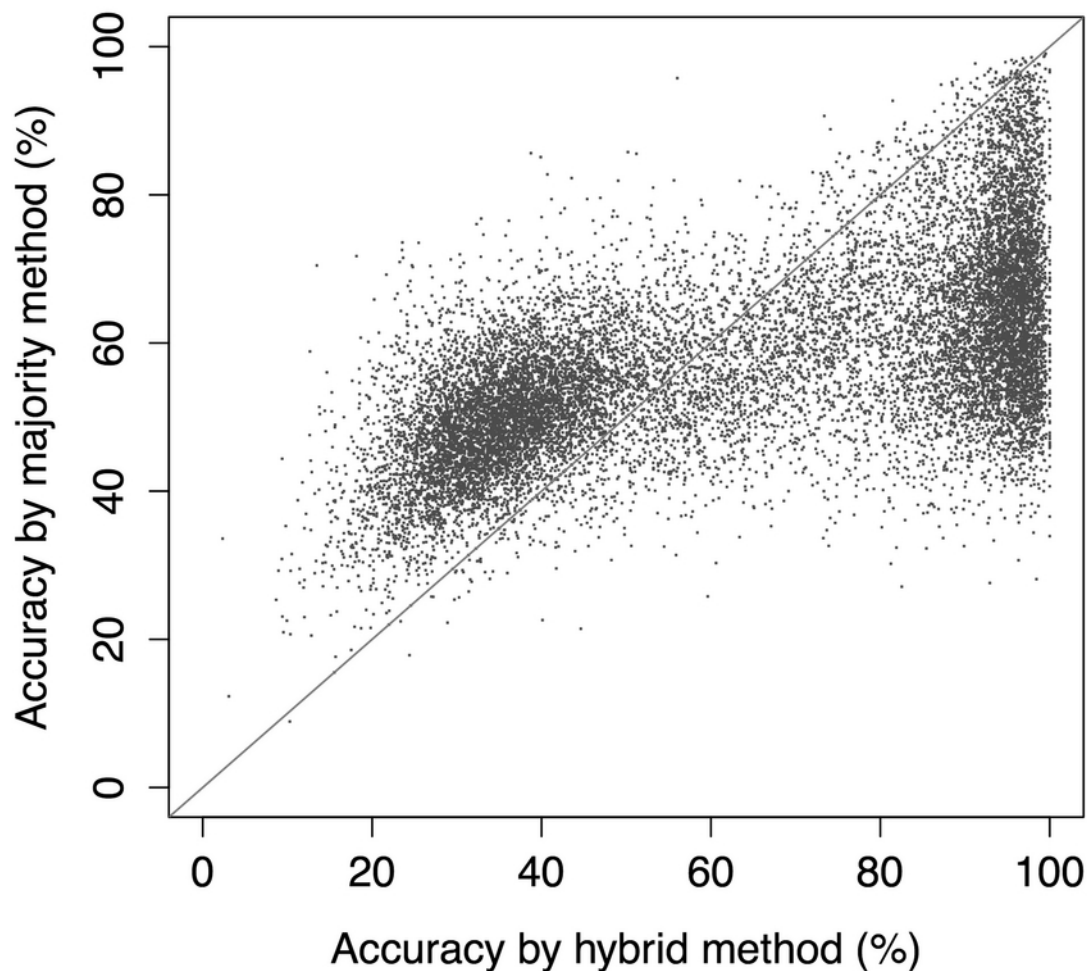
668

669 **Figure 4. Histogram depicting the distribution of the observed prediction accuracies for**

670 **15,544 query proteins by *hybrid method* using *noise filtering scheme*.**

671

672

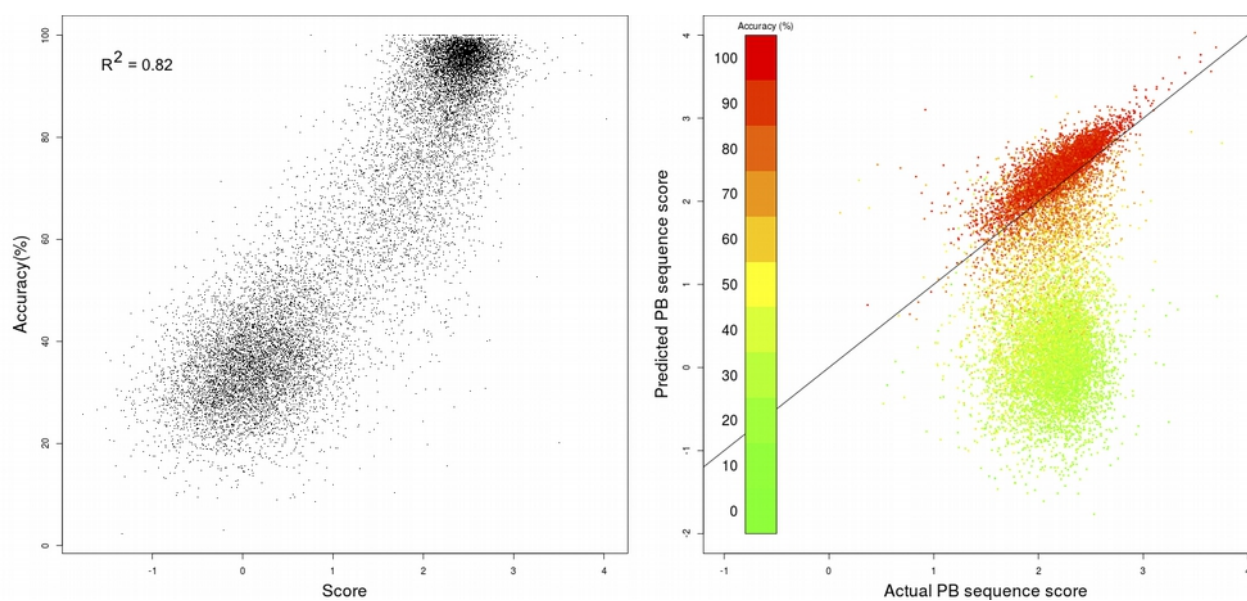


673

674 **Figure 5. Comparison of the majority rule method without *noise filtering scheme* and the**
675 ***Hybrid method with noise filtering scheme*.** Shown are the predictions accuracies for the 15,544
676 query proteins from PDB30 dataset. The diagonal line separates the points where the majority
677 method performs better and the points where the *hybrid method* performs better. Points lying
678 along the diagonal (bisector) represents the situation where both the methods perform equally.

679

680



681

682 **Figure 6. Assessment of the ability of the scoring function to estimate the prediction**
683 **accuracy of PB-kPRED. (a) Scatterplot of score versus accuracy for the 15,544 query proteins**
684 **of PDB30 dataset. (b) Scatterplot of scores for predicted versus actual PB sequences for the**
685 **15,544 query proteins of PDB30 dataset. Datapoints are coloured based on level of accuracy of**
686 **predictions.**

687

688 **Table 1. Content of the different subsets of PDB in terms of pentapeptides accessible to PB-**
689 **kPRED.** Shown here are the total number of pentapeptides and unique pentapeptides for the full
690 PDB and for subsets of PDB filtered at different sequence identity cut-off values.
691

Seq. identity cut-off values (%)	Total number of chains	Total number of unique pentapeptides	Total number of pentapeptides
<30	24,564	1,742,890	5,126,423
<40	28,590	1,881,813	6,153,846
<50	32,588	1,985,203	7,074,571
<70	37,741	2,095,663	8,336,277
<90	42,594	2,148,064	9,351,888
<95	44,714	2,157,239	9,768,748
<100	64,129	2,189,924	14,791,285
Full PDB	274,920	2,268,307	68,621,454

692 **Table 2. Assessment of the richness of PENTAdb towards knowledge-based prediction of**
693 **protein backbone in terms of protein blocks.** Shown is the percentage of pentapeptide queries
694 for which the correct local conformation was found in the list of all possible PBs reported by the
695 PB-kPRED algorithm after querying PENTAdb. A total number of 15,544 query proteins not
696 sharing more than 30% sequence identity (PDB30 dataset) was used in this assessment.

Sequence identity cut-off values (%)	Correct PB found in the list of all possible PBs (%)
<30	71.4%
<40	71.5%
<50	71.6%
<70	71.9%
<90	72.5%
<95	72.8%
<100	77.3%
FULL PDB*	99.93%

697 * all PDB chains (FULL PDB) were accessible for prediction by PB-kPRED but excluding the query
698 protein.

699

700 **Table 3. Evaluation of performance of PB-kPRED knowledge-based approach to predict**
701 **local conformations of protein backbone in terms of protein blocks.** Shown are the accuracies
702 for the PDB30 dataset using both the *majority rule method* and *hybrid method*. For each of the
703 15,544 query protein sequences, the portion of PENTAdb accessible for prediction was
704 dynamically determined using MySQL queries: only pentapeptides coming from protein chains in
705 PDB that shared sequence identities below the indicated cut-off values were accessible to PB-
706 kPRED for prediction of local structures in terms of PBs.

Experiment number	Sequence identity cut-off values (%)	Majority rule method	Hybrid method without noise filtering
A1	FULL PDB*	58.0%±12.9	54.6%±22.0
A2	<100	44.0%±14.7	48.0%±18.7
A3	<95	40.4%±12.7	42.4%±16.1
A4	<90	40.1%±12.5	42.0%±15.9
A5	<70	39.7%±12.5	41.4%±15.8
A6	<50	39.4%±12.5	41.0%±15.9
A7	<40	39.3%±12.5	40.9%±15.9
A8	<30	39.2%±12.5	40.8%±15.9

707

708 * All PDB chains (FULL PDB) were accessible for prediction by PB-kPRED but excluding the
709 query protein.

710

711 **Table 4. Assessment of the performance of PB-kPRED using the hybrid method with *noise***
712 ***filtering scheme* for querying the database.** For each query protein, the portion of the database
713 accessible to the algorithm is first restricted to the closest homologues and if no hits were found,
714 only then the more distant homologues are made accessible progressively. Eight results shown
715 here correspond to the eight experiments represented schematically in Figure 3. Shown are the
716 prediction rates (or Q_{16}) averaged over 15,544 query proteins from the PDB30 dataset that was
717 used in this assessment.

Experiment	Closest homologues to be queried	Average prediction rate or Q_{16} (%)
	first	
B1	100 %	66.31±27.62
B2	<100 %	61.61±24.50
B3	<95 %	61.60±24.49
B4	<90 %	61.59±24.48
B5	<70 %	61.59±24.48
B6	<50 %	61.59±24.47
B7	<40 %	61.58±24.47
B8	<30 %	40.79±15.90

718

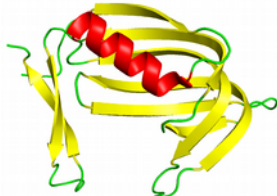
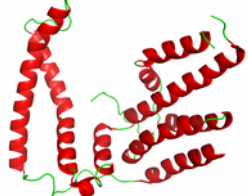
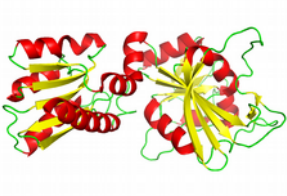
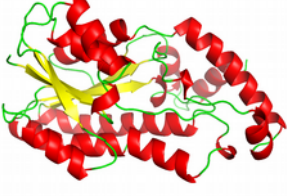
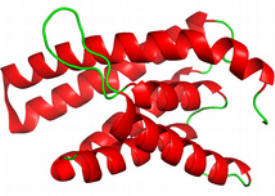
719

720 **Table 5. Assessment of the performance of PB-kPRED and comparison with other**
 721 **previously reported methods.** Shown are the PB-wise prediction accuracies for experiment A8
 722 of the majority method and four different experiments of hybrid method with noise filtering.
 723 These are compared with PB-wise results from LOCUSTRA¹⁹ and the method using Bayesian
 724 approach developed by Etchebest et al²⁵. Experiment B1 PB-wise accuracies were compared with
 725 the other two methods and corresponding cell values in bold represent the best accuracy achieved
 726 between the experiment B1 of *hybrid method*, LOCUSTRA and Bayes method.

PBs	PB frequency	Majority method	Hybrid method						Other methods	
		Expt A8	Expt B8	Expt B4	Expt B2	Expt B1		LOCUSTRA	Bayes method	
		Accuracies					Specificity	MCC	Accuracies	
<i>a</i>	3.68%	34.40%	45.93%	64.60%	64.60%	67.20%	98.15%	0.69	58.16%	56.60%
<i>b</i>	4.29%	15.19%	18.99%	44.52%	44.58%	52.15%	97.72%	0.56	26.14%	20.90%
<i>c</i>	8.31%	23.45%	28.76%	51.62%	51.66%	58.53%	95.95%	0.58	44.81%	32.90%
<i>d</i>	18.68%	39.24%	40.09%	61.83%	61.85%	67.00%	94.12%	0.63	71.58%	54.00%
<i>e</i>	2.18%	21.78%	28.59%	52.30%	52.38%	57.45%	98.96%	0.62	44.74%	38.60%
<i>f</i>	6.45%	24.87%	29.49%	54.64%	54.64%	60.30%	97.21%	0.61	41.45%	30.90%
<i>g</i>	1.14%	10.33%	14.94%	36.79%	36.87%	43.45%	99.19%	0.51	26.84%	30.10%
<i>h</i>	2.10%	24.25%	33.02%	56.90%	56.89%	61.05%	98.82%	0.64	38.45%	40.90%
<i>i</i>	1.49%	20.69%	30.45%	55.36%	55.35%	59.17%	99.18%	0.63	36.87%	38.10%
<i>j</i>	0.83%	15.48%	20.81%	42.57%	42.68%	49.98%	99.40%	0.56	48.19%	49.70%
<i>k</i>	5.25%	30.46%	35.59%	59.30%	59.32%	63.93%	97.67%	0.65	46.46%	33.40%
<i>l</i>	5.20%	26.56%	31.88%	54.87%	54.93%	59.99%	97.69%	0.62	42.71%	35.50%
<i>m</i>	32.89%	60.96%	55.68%	72.38%	72.40%	75.89%	91.03%	0.67	83.76%	70.60%
<i>n</i>	1.78%	26.48%	35.40%	58.16%	58.21%	62.15%	99.05%	0.65	52.08%	50.00%
<i>o</i>	2.44%	29.62%	38.40%	59.49%	59.52%	63.19%	98.70%	0.66	55.10%	48.10%
<i>p</i>	3.29%	23.22%	31.72%	54.27%	54.28%	59.24%	98.25%	0.62	40.80%	29.20%

727

Table 6. Impact of the availability of known homologues on the accuracy of PB-kPRED. Query PDB chains with known homologues and with no known homologues are featured. The *hybrid method with noise filtering* scheme for querying the database was applied for the prediction whereby the conditions were identical to experiment B1 as featured in Figure 3 and Table 6. The queries themselves were excluded from the database prior to prediction. Shown here are the accuracies of the predictions and the numbers of known homologues for different sequence identity thresholds.

Sequence identity thresholds (%)	Queries with known homologues in PDB			Queries with no known homologues in PDB	
	2HX0_A (hypothetical DNA binding protein) 	4HUQ_T (energy-coupling factor transporter EcfT) 	2HXV_A (deminase/ reductase) 	1A27_A (hydroxysteroid dehydrogenase) 	4HZU_S (transmembrane protein associated with Ecf transporter) 
100	2	3	1	1	1
95	2	3	1	1	1
90	2	3	1	1	1
70	2	3	1	1	1
50	2	3	1	1	1
40	2	5	1	1	1
30	2	5	10	1	1
Accuracy (%)	100%	73.41%	39.13%	75.44%	9.37%
Accuracy score	2.81	1.31	0.30	1.99	-0.28