OMICS A Journal of Integrative Biology
Volume 10, Number 1, 2006

Prediction of Protein Functional Specificity without an Alignment

ANDREY FOMENKO, DMITRY FILIMONOV, BORIS SOBOLEV, and VLADIMIR POROIKOV

ABSTRACT

We propose a new approach to predict functional specificity of proteins from their amino acid sequences. Our approach is based on two things: structural Multilevel Neighborhoods of Atom (MNA) descriptors and an original Bayesian algorithm. Usually, a protein sequence is presented as a string of amino acid symbols. Here we introduce a new description of an amino acid sequence: a set of structural MNA descriptors. The MNA descriptor is a string describing an atom and its neighbor atoms according to the selected level. In this work, we also use description of a protein sequence as a set of peptides (strings of amino acid symbols). We performed a case study on two subsubclasses of enzyme nomenclature (EC). It is shown that B-statistics give a sufficient predictive power of enzyme specificity prediction for both MNA descriptors and peptides. We also showed that MNA descriptors give higher accuracy values in comparison with peptides and also provide a choice of MNA descriptor levels for best accuracy prediction. The highest average accuracy prediction that was achieved was 0.98.

INTRODUCTION

A LARGE NUMBER of amino acid sequences (which permanently increases) significantly exceeds the number of proteins with experimentally established functional features. Application of computer methods helps to predict protein function without expensive and time-consuming experiments. There are three basic approaches toward prediction of protein function from amino acid sequence: (1) prediction by homology, (2) pattern or motif searching, (3) functional residue identification.

1. Prediction of protein function by sequence homology involves two steps. One is the detection of homology and other is the inference of function from homology. Homology detection is solved by powerful and sensitive algorithms such as FASTA (Pearson, 2000), BLAST (Altschul and Gish, 1996), PSI-BLAST (Altschul et al., 1997), and HMM (Eddy, 1998). However, the inference of function from homology generates many errors (Rost, 2003; Devos and Valencia, 2001).
2. Functional sites in proteins presented as patterns or motifs are stored in a number of databases such as Prosite (Sigrist et al., 2002), PRINTS (Attwood, 2000), and BLOCKS (Henikoff, 2000). These databases do not provide with successful prediction for many cases. For example Prosite patterns associated with

Laboratory for Structure-Function Based Drug Design, Institute of Biomedical Chemistry, Russian Academy of Medical Science, Moscow, Russia.

the active site of a specific protein are also found in sequences of many unrelated enzymes (Tian et al., 2004).

3. Protein function is determined (in most cases) by several functional residues. Many methods were developed to determine functional residue in protein sequences (Livingstone and Barton, 1993; Casari et al., 1995; Lichtarge et al., 1996; Hannenhalli and Russel, 2000; Landgraf et al., 2001; Mirny and Gelfand, 2002; Rose and Cera, 2002; Del Sol Mesa et al., 2003; Kalinina et al., 2004). These methods detect residues that are not conserved in whole family alignment but are conserved within the small group, which carry the same function. However, as background it is necessary to know about the protein groups in family (Livingstone and Barton, 1993; Hannenhalli and Russel, 2000; Mirny and Gelfand, 2002; Kalinina et al., 2004) or about the three-dimensional (3D) structure of at least one protein per group (Casari et al., 1995; Lichtarge et al., 1996; Landgraf et al., 2001).

All these methods describe an amino acid sequence as a string of symbols, and similarity of sequences is calculated from frequencies of symbols in aligned sequences. We suppose that the same enzyme function can be supported by the structural fragments that are larger or smaller than an amino acid residue or that cannot be described as a consequence of amino acid symbols. We think a more detailed description of a protein chemical structure can give advantage with protein function prediction. Based on this assumption, we introduce the description of a protein sequence as a set of Multilevel Neighborhoods of Atom (MNA) descriptors. The MNA descriptors were developed in our laboratory earlier (Filimonov et al., 1999) and used in the PASS program (Poroikov and Filimonov, 2005) for prediction of biological activity spectra. Earlier we showed that MNA descriptors could be applied for prediction of protein function (Fomenko et al., 2003).

In this paper, we propose a novel approach toward prediction of protein functional specificity that is based on two things: structural MNA descriptors and an original Bayesian algorithm. We suppose that MNA descriptors reflect protein's nature more accurately than amino acid symbols. The predictive algorithm is based on original Bayesian statistics, which is used in the PASS program (Poroikov and Filimonov, 2005) for predicting the biological activity spectra of low-molecular substances. Our approach does not need alignment of sequences in its work, in contrast to all the methods mentioned above.

Our approach supposes the prediction of protein function based on similarity of the query sequence with the different protein groups. So we need a training set to be divided into functional groups. In this work, we performed a case study for our approach using EC.

Enzymes were classified based on their enzyme-catalyzed reactions, and EC was formed by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (Webb, 1992). EC is based on biochemical specificity of enzymes. The EC code includes four classification levels denoted by numbers and separated by points. The first level shows which of the six main classes the enzyme belongs to (1–oxidoreductases, 2–transferases, 3–hydrolases, 4–lyases, 5–isomerases, 6–ligases). The second level indicates the subclass. The second level for oxidoreductases, for example, indicates the group in the hydrogen (or electron) donor, and the second level for hydrolases indicates the nature of the bond hydrolyzed. The third level indicates the sub-subclass. The sub-subclass of oxidoreductases indicates the type of acceptor involved, and the sub-subclass of hydrolases specifies the nature of the substrate. The first three levels of EC code are associated with the chemical reaction, which is catalyzed by the enzymes. For example, 1.2.1.- is the code of oxidoreductases acting on the aldehyde or oxo group of donor with NAD+ or NADP+ as acceptor; 3.4.21.- is the code of serine endopeptidases. The fourth level of EC code is the serial number of the enzyme in its sub-subclass, which represents the substrate specificity of the enzyme reaction.

It is known that many cases of enzymes with the same EC number exist, but which are not homological proteins (Galperin et al., 1998; Babbit, 2003) and vice versa (Babbit, 2003). Thornton et al. (1999) report that about half (91) of the 190-enzyme homologous family in the CATH database (Orengo et al., 1997) have members with different third level EC. This fact caused difficulties with prediction of enzyme function by using EC. However, EC is the largest classification to provide a detailed description of the enzyme function inferred by an experiment.

Understanding this, many researchers investigated the possibility of enzyme function prediction in groups of homological proteins. Todd et al. analyzed sequences from the CATH database aligned by the FASTA algorithm (Pearson, 2000), and came to the conclusion that, even at the low sequence identity of 30%, en-

zyme function could be predicted up to the third EC level with an accuracy of almost 95% (Todd et al., 2001). Wilson et al. (2000) have found that the first three EC levels appear to be conserved down to 40% sequence identity among pairs of sequences with the same fold. Devos and Valencia (2000) showed, using FSSP database (Holm and Sander, 1996), that protein pairs with the same whole EC code were found at 80% sequence identity and higher. At 50–80% identity, only the first three EC levels were identical in the aligned pairs. Below 50% identity, a trend to conservation of the first three EC levels can be observed. Another group of researchers (Tian and Skolnick, 2003) showed that the EC code could be transferred at 60% sequence identity, with accuracy of at least 90%. The same group of authors developed an automatic engine, EFICAz, to infer enzyme function (Tian et al., 2004). This approach allowed 92% accuracy for predicting the fourth EC level in testing sequences, with lower than 40% sequence identity to any member of the training set.

We applied our novel approach to predict enzyme function at the fourth EC level. We used representation of amino acid sequence as sets of MNA descriptors and also as a set of peptides. It is shown that using the Bayesian approach provides high average accuracy values (>0.95) for both MNA descriptors and peptides. However, the use of MNA descriptors is more preferable for this approach, because they give higher values of accuracy than peptides and provide a choice of the descriptor level for the best prediction. Use of MNA descriptors provided an average accuracy of 0.98 for this prediction of enzyme specificity.

METHODS

We formed a training set from two sub-subclasses that belong to the different EC classes oxidoreductases (1.-.-.-) and hydrolases (3.-.-.-) The 3.4.21.- sub-subclass containes serine endopeptidases, which include trypsine and subtilisin families. The 1.2.1.- sub-subclass of oxidoreductases is acting on the aldehyde or oxo group of donor with NAD+ or NADP+ as acceptor. Each sub-subclass contains a large amount of sufficiently diverged proteins. The 35th release of ENZYME database (Bairoch, 2000) was used as a source of the sequences with known EC numbers. The sequences which have more than one EC number and which have non-canonical letters (B, Z, and X) were excluded from the training set. We used only the EC numbers presented by two or more sequences. Sequences, which have the same whole EC code, are defined as a group. The final experimental set contained 77 groups (EC codes) and 1267 sequences.

As one can see in Figure 1, most of the groups consist of two to five sequences, and only four groups contain more than 100 sequences.

In our approach, we compared a query sequence with training set sequences. Training set sequences belonged to different functional groups composed according to the 4th EC level. We assayed similarity between the query sequence and each group in the training set, using Bayesian approach. The original B-statistics (Bayesian statistics) were calculated using the following algorithm.

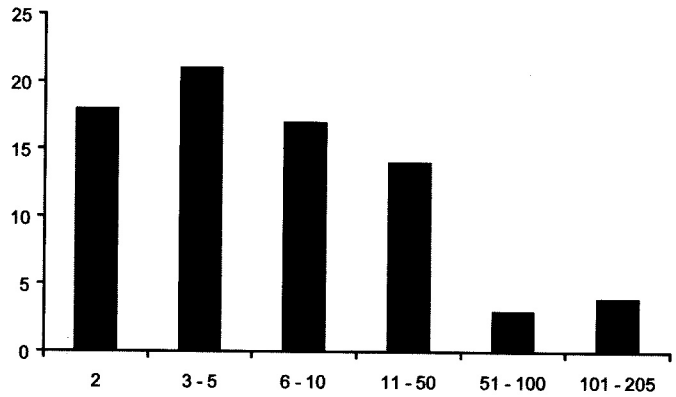

FIG. 1. The representation of the different Enzyme Nomenclature (EC) codes. Number of EC codes (*y*-axis) versus number of sequences enclosed by correspondence interval (*x*-axis).

Let d be a structural element of an amino acid sequence (descriptor). Each amino acid sequence can be described as a set of descriptors d. Let s be a training set sequence. So for each descriptor d of a query sequence we can define $f_s(d)$ as a similarity feature between the query sequence and the training set sequence, if sequence s contain d.

For each group g in training set we calculated two probability estimations. The first one is the conditional probability $P(g \mid d)$ of that a query sequence belongs to a group g assuming that a descriptor d has occurred.

$$P(g|d) = \frac{\sum\limits_{s,s\in g} f_s(d)}{\sum\limits_{s} f_s(d)}$$

Another estimation $P(g)$ is the composite probability of that a query sequence belongs to a group g calculated for all the query sequence descriptors.

$$P(g) = \frac{\sum\limits_{d}\sum\limits_{s,s\in g} f_s(d)}{\sum\limits_{d}\sum\limits_{s} f_s(d)}$$

Then we calculated original B-statistics, which is also described in work (Borodina et al., 2004):

$$S_g = \sin\left[\sum_{s} \arcsin(2P(s \mid d) - 1)/t\right]$$

$$S_{0g} = 2P(g) - 1$$

$$B = (S_g - S_{0g})/(1 - S_g S_{0g})$$

Here, t is the number of unique MNA descriptors in query sequence.

Note that if $P(g \mid d) = 1$ for all d then $B = 1$, if $P(g \mid d) = 0$ for all d then $B = -1$, if $P(g \mid d \approx P(g)$ (there is no relation between the query sequence descriptors and the descriptors of group g) then $B = 0$.

In this work we carried out leave-one-out cross validation procedure for evaluating our approach. According to this procedure, we excluded each sequence from the training set and then used it as a query sequence. So for each protein we calculated B statistics for each group. To assess accuracy of the method used, we calculated independent accuracy prediction (IAP) scores. The IAP score was calculated for each group of the training set by the following way.

Let X be a number of sequences in group g. Let Y be a number of sequences that are not belong to group g. We made up the matrix $X \times Y$. In this matrix we counted $N(B_X > B_Y)$ as number of cells, for which B-statistics of sequence from X is more than B-statistics of sequence from Y. $X \times Y$ is the number of cells in $X \times Y$ matrix.

$$IAP = \frac{N(B_X > B_Y)}{X \times Y}$$

In our approach, each amino acid sequence was represented as a set of MNA descriptors. MNA (Multilevel Neighbourhoods of Atom) descriptor is a string that describes a fragment of the molecular structural formula. The MNA descriptor of the level 0 describes a single atom. The MNA descriptor of the 1st level describes an atom and atoms, which is bound with them by a covalent bond (i.e., neighbors). The MNA descriptor of the 2nd level describes atom and atoms, which distanced from it by one and two covalent bonds. And so on. MNA descriptors of different levels describe molecular fragments of different sizes. Approximately we can say that the descriptor of the 3rd level corresponds to structural fragments from 1 or 2 amino acid residues. The MNA descriptor of the 6th level corresponds to structural fragments that cover 3 or 4 amino acid residues. The MNA descriptor of the 9th level takes the fragments, which cover 5 or 6 amino acid residues. And so forth.

MNA descriptors are build from structural formulas of amino acids that allow comparing structural features of sequences directly in contrast to symbol's description of amino acids. In this work we used descriptors of different levels ranged from one to twelve. To compare our approach with the symbol's representation of sequences, each amino acid sequence was represented also as set of peptides. We used eight levels of peptide sets. Each level was presented by fragments of amino acid string that consist of one, two and so fourth up to eight amino acid symbols in a string.

It can be many different methods to calculate $f_s(d)$ similarity feature. In this work we used two methods for this. The method of descriptor compositions (hereafter *C-method*) did not take into account the descriptor frequencies in amino acid sequence:

$$f_s(d) = 1$$

The method of descriptor frequencies (hereafter *F-method*) took into account the frequency of descriptors in each sequence:

$$f_s(d) = e^{(m-n)\ln(n/m)/2}$$

Here m is the number of descriptor d in query sequence, n is the number of descriptor d in sequence s.

Statistical analysis

So we calculated the set of MNA descriptors of different levels (1–12) for each sequence in the training set. We also formed the set of peptides for each sequence of the training set with different length (from single amino acid to octapeptides). For each group (EC code) we calculated B-statistics for each sequence of training set using leave-one-out cross validation procedure. B-statistics was calculated using both *C-method* and *F-method*. The IAP values were calculated for each group. Finally, 77 IAP values were calculated for both MNA descriptors and peptides type of amino acid description and both *C-method* and *F-method* for all used levels.

RESULTS

Minimal, average, and median values were calculated for IAP values of each MNA and peptide level. These results are presented in Figure 2.

The following results were obtained by using MNA descriptors. Use of *C-method* gives median values more than 0.95 on the 4th to 12th MNA level. Median reaches 1 on the 6th MNA level and holds this value up to the 12th level. Median, observed for *F-method* has values more than 0.95 at all the MNA levels. It takes on 0.985 value on the 1st MNA level then it grows up to 1 at the 6th MNA level and holds this value up to the 12th MNA level. Average of IAP that was observed for *C-method*, has value more than 0.95 on the 5th to 12th MNA levels. The highest average value is 0.986, which is observed on the 10th MNA level. For *F-method* more than 0.95 average values are observed on 6th to 12th MNA level. The highest average value is 0.982 on the 10th MNA level. The highest minimal value for *C-method* is 0.487, which is observed on the 7th MNA level and for *F-method* is 0.08 for the 10th level.

The following results were obtained for peptides. Median, obtained with *C-method*, has values close to 1 on the 2nd to 4th peptide level and median, obtained with *F-method* has values close to 1 on the 1st to 4th peptide level. For both methods the highest median values is 0.999 on the 3rd peptide level. Average, observed for both methods, has value more than 0.95 on the 3rd and 4th level of peptide length. The highest average value is 0.963 on the 3rd peptide level for both methods. The highest minimal value for both methods is 0.385, which observed on the 4th peptide level.

It is obvious that one can choose MNA or peptide level with satisfactory predictive power for each of four used variants of predicting (MNA with *C-method*, MNA with *F-method*, peptides with *C-method*, peptides with *F-method*). So the B-statistics shows good results of the enzyme specificity prediction for both peptides and MNA-descriptors were used. However median and average IAP values are higher for MNA descriptors. Thus the use of MNA descriptors gives more accurate predictions than the use of peptides.

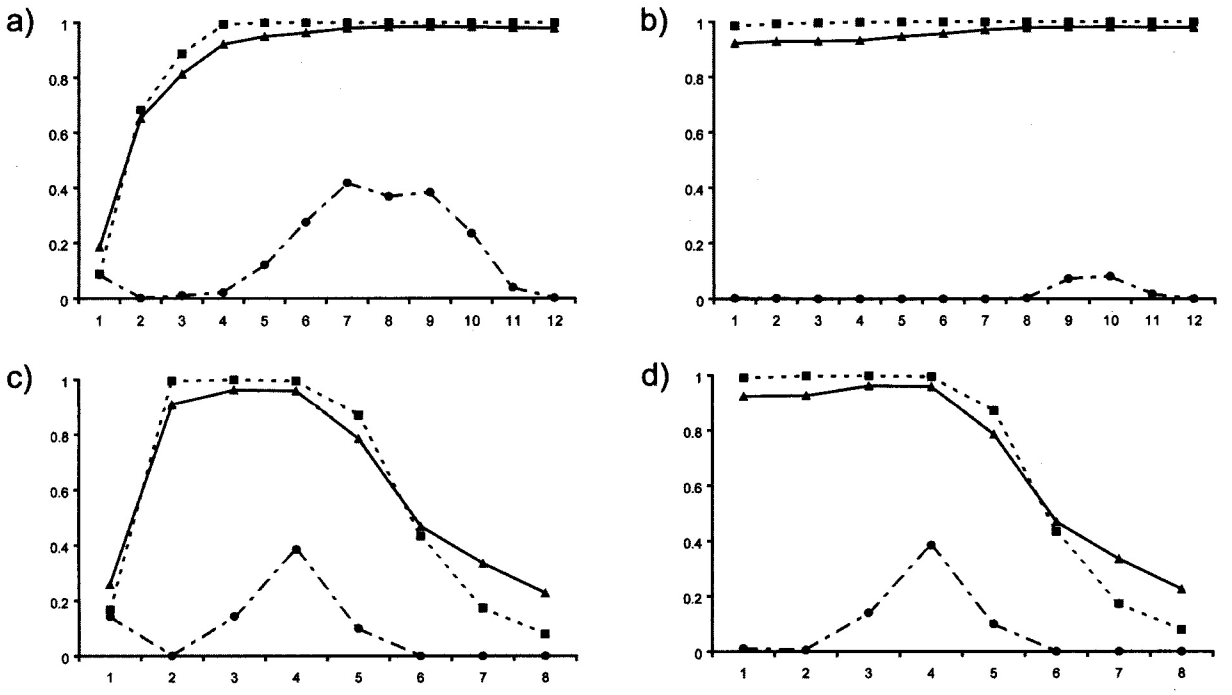

FIG. 2. (**a,b**) Independent accuracy prediction (IAP) values obtained for MNA descriptors. (**c,d**) IAP values obtained for peptides. (**a,c**) IAP values obtained with using C-method. (**b,d**) IAP values obtained with using F-method. The level of the correspondent Multilevel Neighborhoods of Atom (MNA) descriptors or peptides is shown on *x*-axis; IAP value is on *y*-axis. Average values are shown by triangles (firm line), median values are shown by squares (dotted line), and minimal values are shown by squares (dashed line).

It is clear that both methods (*C-* and *F-methods*) have the best accuracy on the same levels of MNA descriptors (8th to 12th) and on the 3rd level of peptide. It says that using of frequencies of MNA descriptor in sequence do not give any advances in accuracy values. Nevertheless *F-method* gives satisfactory IAP value (>0.95) at the lower levels of MNA descriptors or peptides. It can give advantage with time of calculations. For example, calculations for the 2nd MNA level is more than twenty times quickly then for the 12th MNA level. This condition could be sufficient for the bigger size of training set.

It can be seen in Figure 2 that MNA descriptors up to the 12th level hold high accuracy values, whereas IAP values for peptides decrease quickly after the 4th peptide level. Another representation of the results in figure 3 can help to clarify this observation and give one more reason for progress of MNA descriptors.

The use of MNA descriptors reveals large area (six MNA levels: from the 8th to the 12th) with only four groups (1.2.1.1, 1.2.1.16, 1.2.1.22, 1.2.1.28) having IAP lower than 0.95 for both C-method and F-method used (Fig. 3). On the contrary, the light area obtained for peptides is smaller. In fact, only the 4th peptide level has the smallest number of dark fields (eight dark fields for *C-method* and seven dark fields for *F-method*). This observation confirms the conclusion about the accuracy progress of the results obtained with MNA descriptors. Moreover, results from Figure 3 are consequence of the nature of the descriptions used for the amino acid sequences. One level of MNA descriptor increased by an atom for each side of molecular chain, whereas one level of peptides increased by one amino acid residue with many atoms. Indeed, one level of peptides corresponds to more than one MNA levels. For example, all the MNA descriptors of the 1st, the 2nd, and the 3rd levels are found within a structural formula of a single amino acid residue or they are the same for all amino acid residues (except the descriptors of some atoms of proline and glycine). So nature of MNA descriptors is reflected by more smooth changes of IAP values. This property of MNA descriptors could be used to choose the descriptors level with the best prediction accuracy for the interesting specificity group.

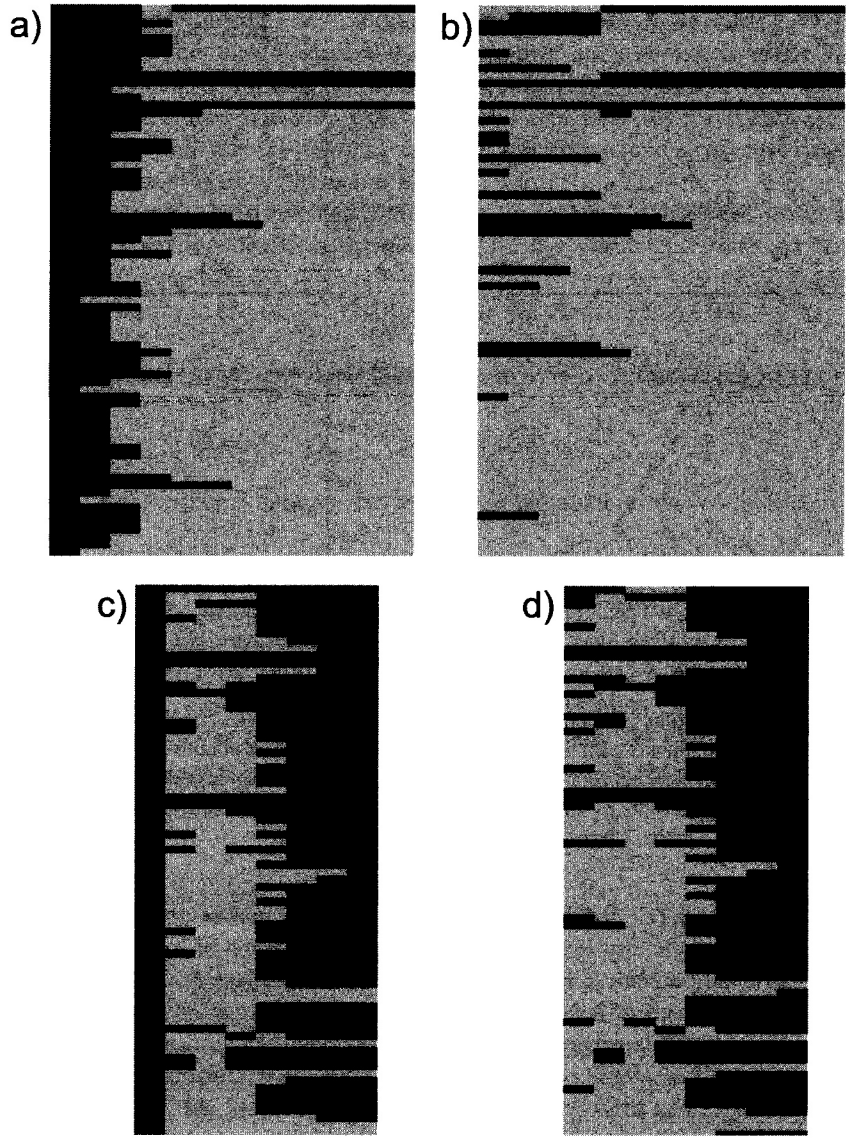

FIG. 3. There are two kinds of independent accuracy prediction (IAP) values in this figure. Dark fields show groups with IAP values lower than 0.95; light fields show groups with IAP equal or higher than 0.95. Level of Multilevel Neighborhoods of Atom (MNA) descriptors or peptides grows up from left to right. (**a**) Obtained for MNA descriptors using C-method. (**b**) Obtained for MNA descriptors using F-method. (**c**) Obtained for peptides using C-method. (**d**) Obtained for peptides using F-method.

DISCUSSION

In this work, we introduce a novel approach toward prediction of protein functional specificity. Our approach is based on B-statistics and structural MNA descriptors. We used two representations of amino acid sequence as a set of MNA descriptors and a set of peptides. We showed that B-statistics give a high average accuracy value for both peptides and MNA descriptors. However, MNA descriptors showed higher accuracy values than peptides. Also, MNA descriptors provided a choice of descriptor levels for the best prediction. So, the use of MNA descriptors is more preferable than the use of peptides in this approach. The highest average accuracy that was achieved with MNA descriptors in this work was 0.98.

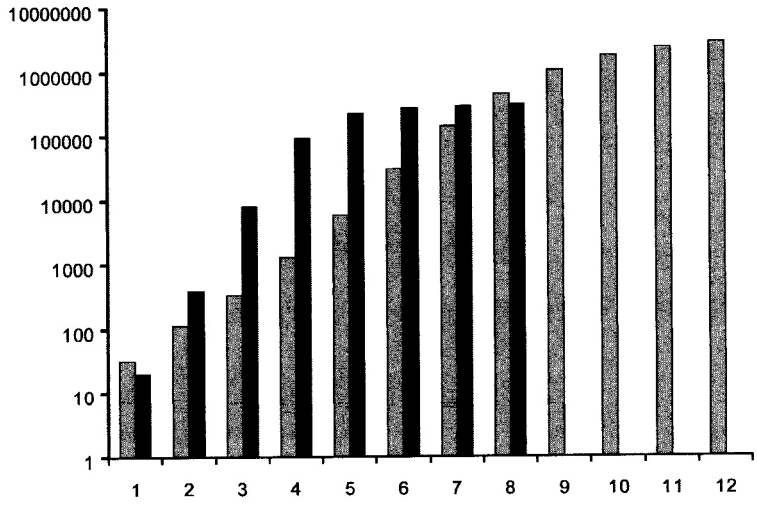

FIG. 4. Number of unique elements (y-axis) at each level of used amino acid description. The level of Multilevel Neighborhoods of Atom (MNA) descriptors (light) or peptides (dark) is shown by x-axis.

We think that MNA descriptor advances could be explained after analysis of a number of unique MNA descriptors and peptides in the dataset. As we noted previously, MNA descriptor of the 6th level approximately corresponds to the 3rd or the 4th level of peptides and MNA descriptor of the 9th level corresponds to the 5th or the 6th peptide level. Taking this rule into account, we can say that the number of unique elements in MNA descriptors is more than the number of unique elements of peptides for the 8th to the 12th levels, which take the highest IAP values with MNA descriptors (Fig. 4). Furthermore, MNA descriptors are calculated for each atom. So number of unique elements of MNA descriptors tends to number of atoms in amino acid sequence with growth of MNA level. Apparently, larger number of unique elements of MNA descriptors, especially for the high MNA levels, gives advantage in predictive power.

Of course, this work performed only with two EC subsubclasses (1.2.1.- and 3.4.21.-), which are represented by 1267 amino acid sequences. The 35th release of ENZYME database would give more than 48,000 sequences. However, the size of the dataset used is near the maximum of our program's possibility. In the future, we plan to optimize the program algorithm for the larger dataset that would be used. On the other hand, if the dataset were larger, we think that the accuracy of prediction would decrease. Nevertheless, we suppose that the biggest part of EC numbers will show sufficiently high accuracy of prediction. So we plan to develop this approach toward an expert system for the prediction of enzyme specificity.

Although we showed our approach only in the context of enzyme specificity prediction we presented it as a more general prediction of protein functional specificity. Thus, we propose that our approach could be applied for other types of functional prediction such as prediction of type of enzyme reaction (EC subsubclasses) or prediction of function of non-enzymes. Also we hope that this approach could be applied for tasks of structure prediction such as fold prediction or secondary structure prediction. We plan to try another sort of prediction in the future.

CONCLUSION

We introduced a new approach toward predicting of protein functional specificity based on MNA descriptors and original B-statistics used in the PASS program (Poroikov and Filimonov, 2005). We tested our approach for the task of enzyme specificity prediction. Usually amino acid sequence is described as sequence of symbols for the task of functional prediction. Here we applied new description of amino acid sequence as set of structural MNA descriptors. In this work we used two types of description of amino acid sequence: the one is a set of MNA descriptors and the other is a set of peptides. It is shown that the use of

B-statistics provides sufficient predictive power for both types of protein sequence description. However, MNA descriptors shows higher accuracy than peptides and provides a possibility to choose a descriptors level with the best accuracy value. The highest average accuracy value achieved was 0.98. Advances of MNA descriptors could be explained by the nature of MNA descriptors that reflects chemical structure of proteins more accurately than amino acid symbols.

ACKNOWLEDGMENT

This work was supported by the Russian Foundation of Basic Research (grant N 04-04-49390).

REFERENCES

ALTSCHUL, S.F., and GISH, W. (1996). Local alignment statistics. Methods Enzymol **266,** 460–480.

ALTSCHUL, S.F., MADDEN, T.L., SCHAFFER, A.A., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25,** 3389–3402.

ATTWOOD, T.K., CRONING, M.D., FLOWER, D.R., et al. (2000). PRINTS-S: the database formerly known as PRINTS. Nucleic Acids Res **28,** 225–227.

BABBITT, P.C. (2003). Definitions of enzyme function for the structural genomics era. Curr Opin Chem Biol **7,** 230–237.

BAIROCH, A. (2000). The ENZYME database in 2000. Nucleic Acids Res **28,** 304–305.

BORODINA, Y., RUDIK, A., FILIMONOV, D., et al. (2004). A new statistical approach to predicting aromatic hydroxylation sites. Comparison with model-based approaches. J Chem Inf Comput Sci **44,** 1998–2009.

CASARI, G., SANDER, C., and VALENCIA, A. (1995). A method to predict functional residues in proteins. Nat Struct Biol **2,** 171–178.

DEL SOL MESA, A., PAZOS, F., and VALENCIA, A. (2003). Automatic methods for predicting functionally important residues. J Mol Biol **326,** 1289–1302.

DEVOS, D., and VALENCIA, A. (2000). Practical limits of functional prediction. Proteins **41,** 98–107.

DEVOS, D., and VALENCIA, A. (2001). Intrinsic errors in genome annotation. Trends Genet **17,** 429–431.

EDDY, S.R. (1998). Profile hidden Markov models. Bioinformatics **14,** 755–763.

FILIMONOV, D., POROIKOV, V., BORODINA, Y., et al. (1999). Chemical Similarity Assessment through multilevel neighborhoods of atoms: definition and comparison with the other descriptors. J Chem Inf Comput Sci **39,** 666–670.

FOMENKO, A.E., SOBOLEV, B.N., FILIMONOV, D.A., et al. (2003). Use of structural MNA descriptors for designing profiles of protein families. Biofizika **48,** 595–605.

GALPERIN, M.Y., WALKER, D.R., and KOONIN, E.V. (1998). Analogous enzymes: independent inventions in enzyme evolution. Genome Res **8,** 779–790.

HANNENHALLI, S.S., and RUSSEL, R.B. (2000). Analysis and prediction of functional sub-types from protein sequence alignment. J Mol Biol **303,** 61–76.

HENIKOFF, J.G., GREENE, E.A., PIETROKOVSKI, S., et al. (2000). Increased coverage of protein families with the blocks database servers. Nucleic Acids Res **28,** 228–230.

HOLM, L., and SANDER, C. (1996). Mapping the protein universe. Science **273,** 595–602.

KALININA, O.V., MIRONOV, A.A., GELFAND, M.S., et al. (2004). Automated selections of positions determining-functional specificity of proteins by comarative analysis of orthologous groups in protein families. Protein Sci **13,** 443–456.

LANDGRAF, R., XENARIOS, I., and EISENBERG, D. (2001). Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J Mol Biol **307,** 1487–1502.

LICHTARGE, O., BOURNE, H.R., and COHEN, F.E. (1996). An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol **257,** 342–358.

LIVINGSTONE, C.D., and BARTON, G.J. (1993). Protein sequence alignment: a strategy for the hierarchical analysis of residue conservation. Comput Appl Biosci **9,** 745–756.

MIRNY, L.A., and GELFAND, M.S. (2002). Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors. J Mol Biol **321,** 7–20.

ORENGO, C.A., MICHIE, A.D., JONES, S., et al. (1997). CATH—a hierarchic classification of protein domain structures. Structure **15,** 1093–1108.

PEARSON, W.R. (2000). Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol **132,** 185–219.

POROIKOV, V., and FILIMONOV, D. (2005). PASS: prediction of biological activity spectra for substances. In *Predictive Toxicology.* C. Helma, eds. (Marcel Dekker, New York, pp. 459–478.

ROSE, T., and DI CERA, E. (2002). Substrate recognition drives the evolution of serine proteases. J Biol Chem **277,** 19243–19246.

ROST, B., LIU, J., and NAIR, R., et al. (2003). Automatic prediction of protein function. Cell Mol Life Sci **60,** 2637–2650.

SIGRIST, C.J., CERUTTI, L., HULO, N., et al. (2002). PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform **3,** 265–274.

THORNTON, J.M., ORENGO, C.A., TODD, A.E., et al. (1999). Protein folds, functions and evolution. J Mol Biol **293,** 333–342.

TIAN, W., and SKOLNICK, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol **333,** 863–882.

TIAN, W., ARAKAKI, A.K., and SKOLNICK, J. (2004). EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. Nucleic Acids Res **32,** 6226–6239.

TODD, A.E., ORENGO, C.A., and THORNTON J.M. (2001). Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol **307,** 1113–1143.

WEBB, E.C. (1992). *Enzyme Nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology* (Academic Press, New York).

WILSON, C.A., KREYCHMAN, J., and GERSTEIN, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. J Mol Biol **297,** 233–249.

Address reprint requests to:
Dr. Vladimir Poroikov
Laboratory for Structure-Function Based Drug Design
Institute of Biomedical Chemistry
Russian Academy of Medical Science
10 Pogodinskaya str.
Moscow, 119121, Russia

E-mail: Vladimir.Poroikov@ibmc.msk.ru