Volume 297, number 3, 241–246 FEBS 10690 February 1992

Applying machine learning methods for finding significant amino acid properties in proteins

Joachim Selbig, Frieder Kaden and Ina Koch

Central Institute of Cybernetics and Information Processes, Kurstrasse 33, D-O-1086 Berlin, Germany

Received 9 December 1991

There are several possibilities for definition and derivation of sequence patterns associated with structural motifs, in particular on the secondary structure level which may be used to predict these structure elements. Sequence patterns consist of a number of consecutive positions along the polypeptide chain from which a certain quantity is specified. One of the important factors in deriving sequence patterns in terms of amino acid properties is how to find the most characteristic properties to specify a certain position and thus to avoid redundant physical information. We have applied machine learning methods to select the most significant amino acid properties describing a structurally determined sequence position. Results are given for the beginning of α-helices. These methods may link the gap between amino acid patterns and property patterns and thus are valuable to improve protein structure prediction.

Amino acid properties; Machine learning; Sequence pattern; Structure prediction

1. INTRODUCTION

Understanding the relation between amino acid sequence and local structure in proteins is a fundamental problem in biochemistry. It is a necessary step towards the prediction of the 3D-structure from the sequence. There are several possibilities to define and to derive sequence patterns associated with structural motifs, in particular on the secondary structure level. Such patterns describe protein structure locally and can be used to predict these structure elements. Sequence patterns consist of a number of consecutive positions along the polypeptide chain from which a certain quantity is specified. Basically, amino acid patterns are distinguished from property patterns [1]. Positions in amino acid patterns are specified by mathematical expressions over the set of the 20 naturally occurring amino acids whereas position specifications in property patterns consist of expressions over amino acid properties as defined e.g. in [2]. The 20 amino acids have different physicochemical and biochemical properties such that one and the same segment of the polypeptide chain can be described by several property patterns. Therefore, one of the important factors in deriving sequence patterns in terms of amino acid properties is how to find the most characteristic properties to specify a certain position and thus to avoid redundant physical information.

From the set of the 20 amino acids 2^{20} = 1,048,576 subsets may be derived which may be related to properties or combinations of properties. It is unlikely that all these subsets and the corresponding property combinations are full of meaning for the structure description. But the 10 amino acid classes used in [3] cannot be regarded as sufficient to characterize all the different functionally and structurally determined sequence positions. 71 amino acid classes were defined in [4] from which 48 were used in patterns for structure prediction. The property combinations related to these classes consist of elementary properties such as hydrophobicity, charge, and size connected by the logical operators AND, OR, and NOT. The classes are hierarchically ordered which enables the automatic modification of position specifications by specialization and/or generalization. From this work it follows that structurally meaningful properties are often not explicit and occur jumbled together with other properties. Therefore, there are good reasons to apply methods which can handle and assimilate information found in databases automatically.

Correspondence address: J. Selbig, G.M.D., Institute of Foundations of Information Technology, P.O. Box 1316, D-5205 Sankt Augustin 1, Germany, Fax: (49) (2241) 14 2889.

2. MATERIALS AND METHODS

We use machine learning methods to find significant amino acid properties describing structurally determined sequence positions. In particular, decision trees are learned from the occurrences of the amino acids observed on structurally defined sequence positions in proteins of the Brookhaven Protein Data Bank [5].

The problem of finding relevant amino acid properties can be defined as an identification task a specific type of classification tasks. The learning set consists of the set of occurrences of the 20 amino acids on structurally defined sequence positions. From a given set of properties those are searched for which allow the best identification of each of the 20 amino acid types via these properties or attributes. Critical

for the selection of the properties are the frequencies of the amino acids in the learning set which convey the meaning or the weight of the properties describing them.

For solving such identification and more general classification tasks decision tree algorithms were developed in the field of learning from examples. Decision trees are classification rules which consist of a root node, interior nodes, branches, and leaf nodes (for more details see [6,7]. The interior nodes are tests applied to instances during the classification. Branches from a non-leaf node correspond to the possible test outcomes. The general procedure to learn decision trees from examples is based on successive subdivisions of the learning set. This process aims to discover sizeable subsets of the learning set that belong to the same class. Before introducing a new test, the merit of all attributes with respect to the subdivision process is measured by the mutual information between the classes and the attributes. The best attribute according to this measure is placed as the next test in the tree.

Starting point of our investigations are the results of our protein structure analysis based on methods from applied graph theory [8]. In this approach the data about the spatial structure of the proteins determined by X-ray crystallography are transformed into a graph description, which provides the possibility to define a multitude of amino acid patterns for describing secondary und supersecondary structure elements. These patterns contain information about the pairwise amino acid dependencies within structurally determined environments. Different patterns are used to characterize the beginning, the continuation, and the end of structure elements. According to the sequential distance of the segments specified by the pattern framework we distinguish between short-range and long-range patterns. The consideration of the amino acid dependencies rests upon the computation of distances between atoms of different amino acids. If a preset threshold criterion is satisfied we speak of contacts between the considered amino acids. In dependency of the atom type (backbone atoms, side chain atoms) we distinguish different contact kinds. If the preset threshold criterion is not satisfied in the neighbourhood of contacts we speak of forbidden contacts.

4. RESULTS AND DISCUSSION

From a set of 84 proteins* 67 patterns were derived with the method outlined in [8] from which the patterns H_1, H_2, H_6, and H_7 describing the beginning of α-helices are considered more into detail (see Fig. 1). Each of the patterns covers 6 sequence positions and is characterized by the values of some parameters (see Table I). Among others, these parameters describe the number and the type of contacts between the amino acids and the sequential distance between the amino acids forming the main contact. By projection of the pairwise amino acid dependencies onto the pattern positions we find the corresponding sets of amino acid occurrences. Figs. 2 and 3 show the distributions of the amino acids on the positions i, i+3 and i+4, and i, i+1 and i+4, respectively, of the two types of the beginning of α-helices. To find the most important amino acid properties for a specific sequence position by decision tree methods a certain set of properties must be given. Table II presents the used set of 16 properties [2,9]. In further investigations other properties will be selected from the large set of more than 200 possible ones [10].

Figs. 4 and 5 show the decision trees for the position i of type 1 and position i+4 of type 2, respectively, related to the amino acid occurrences of the corresponding patterns. The numbers characterizing the test at-

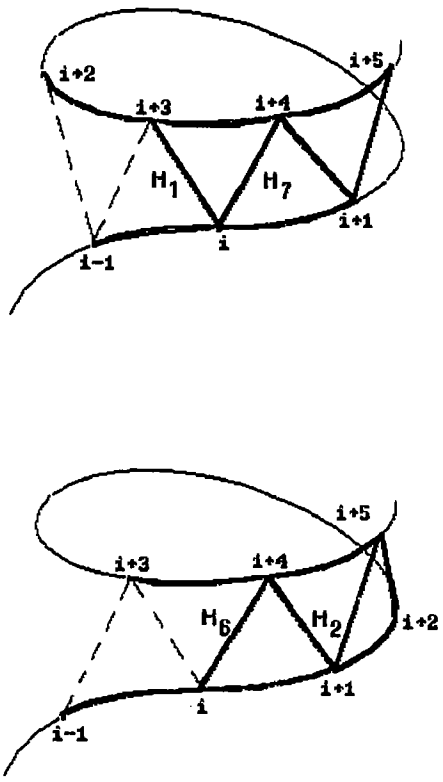

Fig. 1. Two types of the beginning of α-helices. Dashed lines characterize forbidden contacts. (a) Type 1 is described by the two patterns H_1 and H_7 for which 381 and 452 instances, respectively, were observed. Main contacts exist between the amino acids i and i+3 (H_1) and i and i+4 (H_7), respectively. (b) Type 2 is characterized by the two patterns H_2 and H_6 for which 76 and 53 instances, respectively, were observed. Main contacts exist between the amino acids i+1 and i+4 (H_2), and i and i+4 (H_6), respectively.

Table 1

Sequence positions covered by the 4 considered patterns (see Fig. 1)

	1	2	3	4	5	6
H_1	i-1	i	i + 1	i + 2	i + 3	i + 4
H_2	i	i + 1	i + 2	i + 3	i + 4	i + 5
H_6	i-1	i	i + 1	i + 3	i + 4	i + 5
H_7	i-1	i	i + 1	i + 3	i + 4	i + 5

*Corresponding to the Brookhaven Protein Data Bank entries [5]: 156B, 1ABP, 1BP2, 1CAC, 1CC5, 1CCR, 1CHG, 1CTF, 1CY3, 1CYC, 1ECA, 1FB4, 1FDH, 1FDX, 1FX1, 1GCR, 1GP1, 1HDS, 1HHO, 1HIP, 1HMQ, 1LZ1, 1MBS, 1NTP, 1PP2, 1PPD, 1REI, 1RHD, 1RN3, 1RNS, 1RNT, 1SBT, 1SGC, 1TGB, 1TGS, 1TGT, 1TIM, 1TON, 1TPA, 1UBQ, 2ABX, 2ACT, 2ALP, 2APP, 2AZA, 2CAB, 2CCY, 2CDV, 2CPP, 2CTS, 2CYP, 2EBX, 2FD1, 2GCH, 2GN5, 2GRS, 2KAI, 2LH1, 2LZM, 2OVO, 2PAB, 2RHE, 2SNS, 3APR, 3C2C, 3CPV, 3EST, 3RP2, 3TPI, 451C, 4APE, 4DFR, 4HHB, 4LDH, 4LYZ, 4PTI, 5CPA, 5RSA, 5RXN, 6CHA, 6PCY, 7CAT, 7TLN, 9PAP.

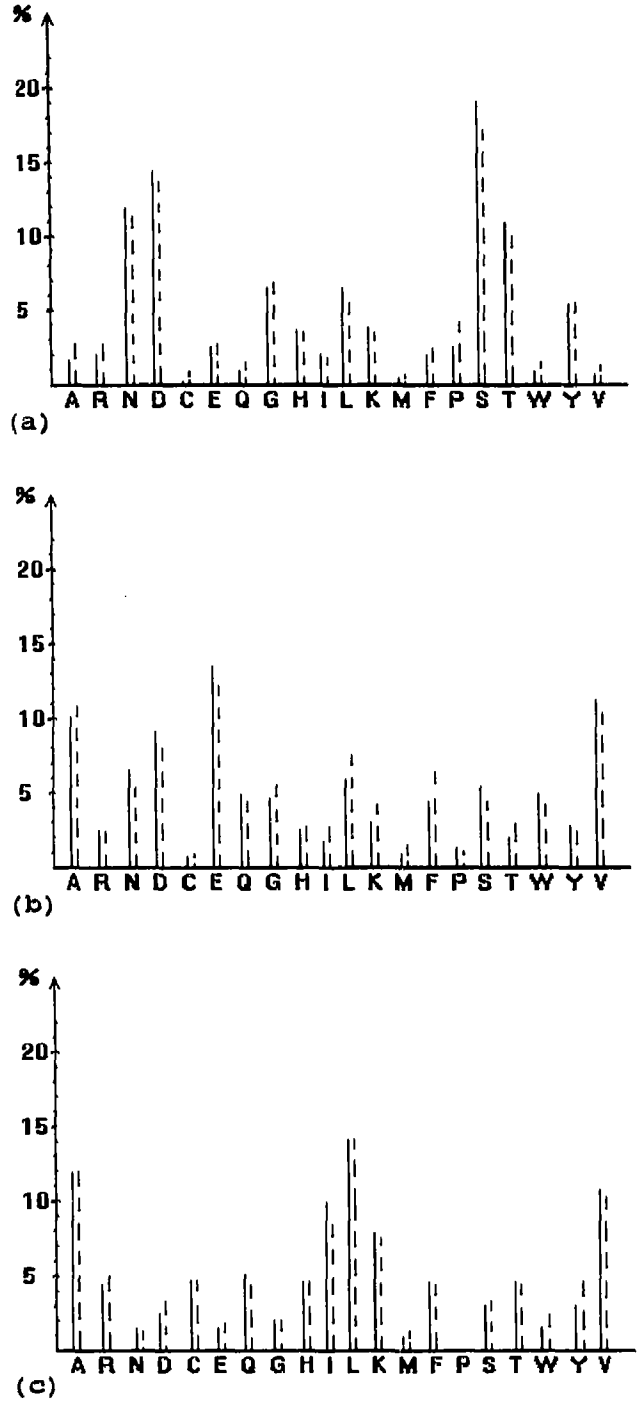

Fig. 2. Distributions of the amino acids occurring in the patterns H_1 (traced) and H_7 (dashed) of type 1. Amino acids are given in the one-letter code. (a) Position *i*, (b) Position *i*+3, and (c) Position *i*+4.

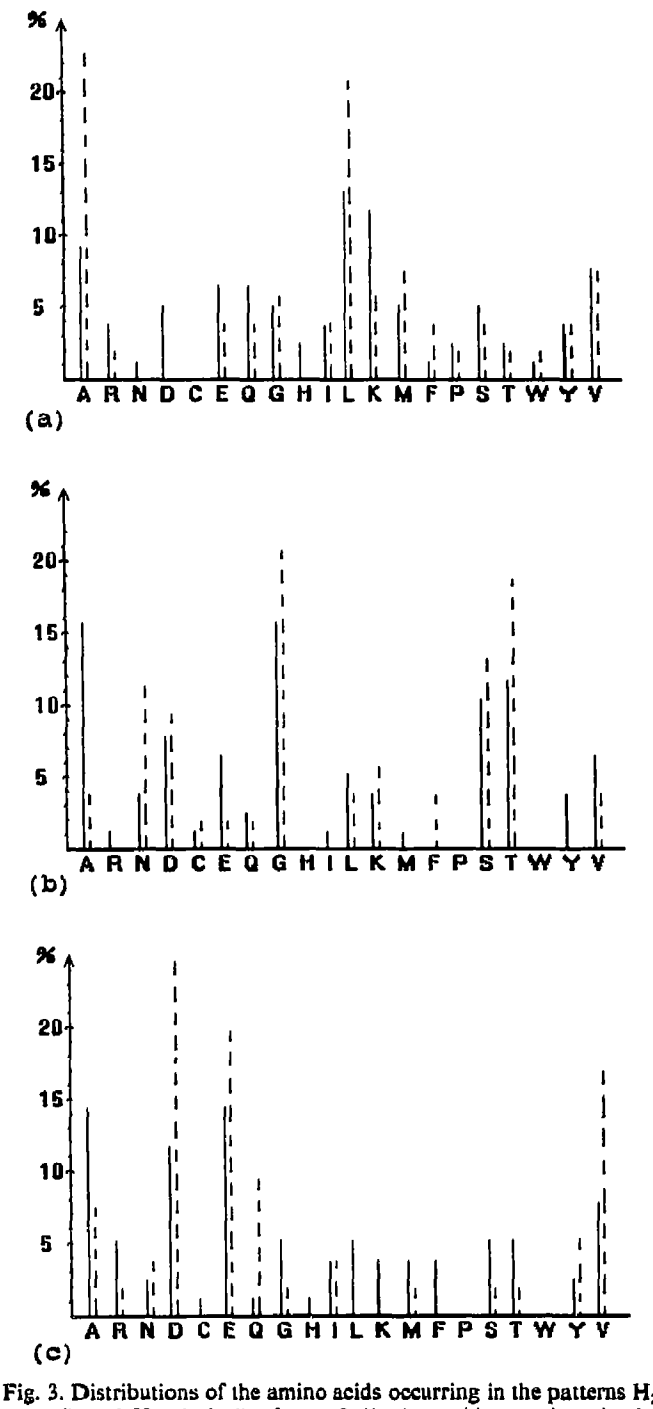

Fig. 3. Distributions of the amino acids occurring in the patterns H_2 (traced) and H_6 (dashed) of type 2. Amino acids are given in the one-letter code. (a) Position *i*, (b) Position *i*+1, and (c) Position *i*+4.

tributes correspond to the amino acid properties of Table II. The distributions in Fig. 2 related to the patterns of type 1 are very similar. This is reflected by the corresponding decision trees in that only with few exceptions the same properties are used on the same positions as test attributes and thus to characterize the amino acids (see Fig. 4a and 4b). In particular, the beginning of α-helices of type 1 is characterized by the fact that no proline was observed on position *i*+4.

The distributions in Fig. 3 related to the patterns of type 2 are very different which is reflected by the corresponding decision trees in Figs. 5a and 5b. They have different structures and contain different properties for

Table II

Used amino acid properties.

The amino acids are given in the one-letter code. '1' in position (i, j) of the matrix means that amino acid j has property i and '0' that it doesn't have this property.

		A	R	N	D	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	hydrophobic	1	0	0	0	1	0	0	1	1	1	1	1	1	1	0	0	0	1	1	1
2	polar	0	1	1	1	0	1	1	0	1	0	0	1	0	0	0	1	1	1	1	0
3	negativ charged	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	positiv charged	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
5	small	1	0	1	1	1	0	0	1	0	0	0	0	0	0	1	1	1	0	0	1
6	tiny	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0
7	aliphatic	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1
8	aromatic	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	1	0
9	side chain with <4 heavy atoms	1	0	0	0	1	0	0	1	0	0	0	0	0	0	1	1	1	0	0	1
10	side chain with >4 heavy atoms	0	1	0	0	0	1	1	0	1	0	0	1	0	1	0	0	0	1	1	0
11	linear side chain	1	0	0	0	1	0	0	1	0	0	0	1	1	0	1	0	1	0	0	0
12	bulky side chain	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	0	1	1	1	1
13	side chain with O (oxygen)	0	0	1	1	0	1	1	0	0	0	0	0	0	0	0	1	1	0	1	0
14	side chain without NH, OH, SH	1	0	0	0	0	0	0	1	0	1	1	0	1	1	1	0	0	0	0	1
15	charged side chain	0	1	0	1	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0
16	side chain with <3 CH	0	0	1	1	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0

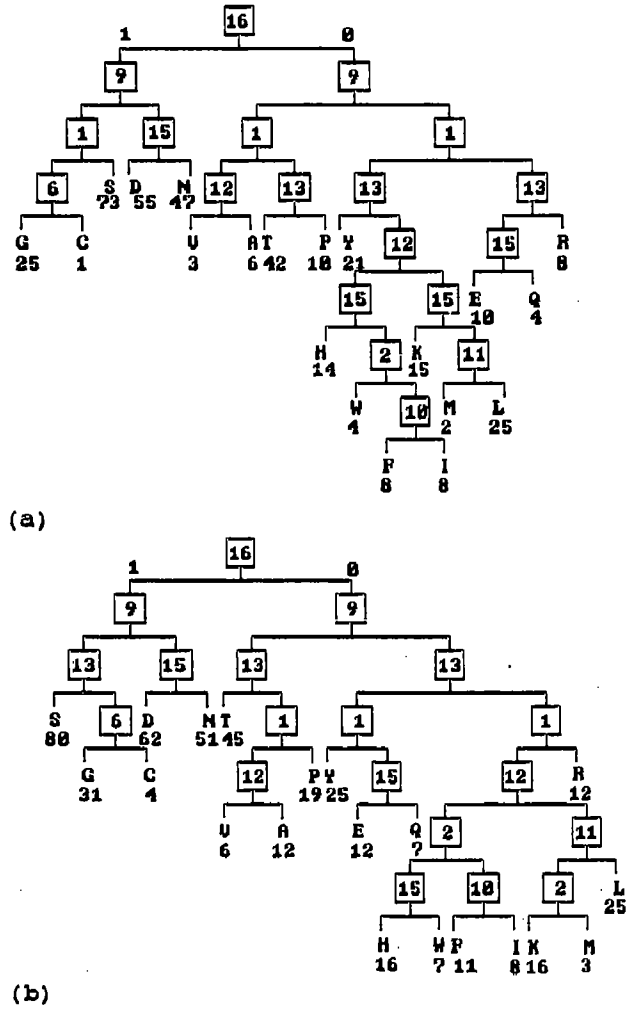

Fig. 4. Decision trees for position i of the main contacts of type 1. (a) Pattern H_1. (b) Pattern H_7.

characterizing the amino acids. In particular, the beginning of α-helices of type 2 is characterized by the fact that no proline and no tryptophan were observed in positions i+1 and i+4, and no cysteine in position i, and

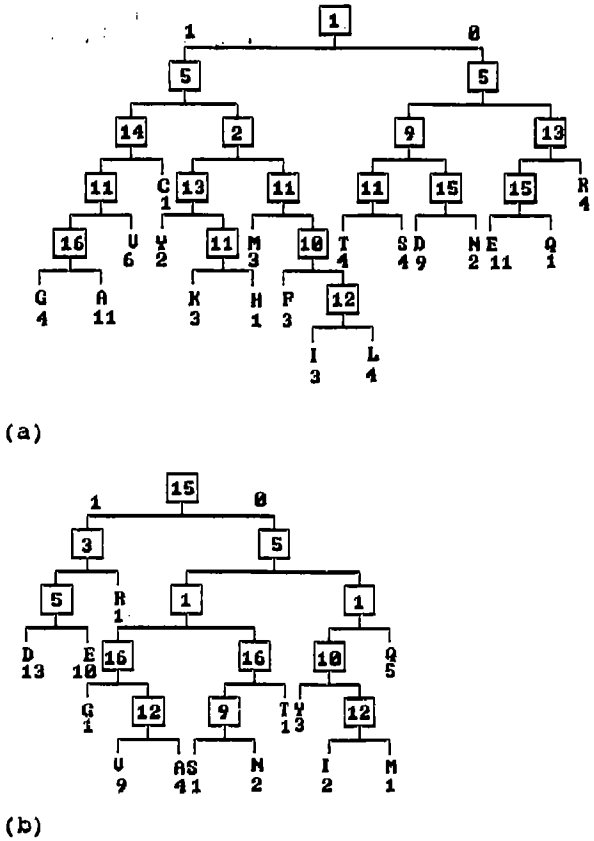

Fig. 5. Decision trees for position i+4 of the main contacts of type 2. (a) Pattern H_2. (b) Pattern H_6.

```
If    (hydrophobic)
then  if    (small)
      then  if    (side chain without NH, OH, SH)
            then  if    (linear side chain)
                  then  if    (side chain with < 3 CH)
                        then  (C)    else  (A)
                  else  (V)
            else  (C)
      else  if    (polar)
            then  if    (side chain with O)
                  then  (Y)
                  else  if    (linear side chain)
                        then  (K)    else  (H)
            else  if    (linear side chain)
                  then  (M)
                  else  if    (side chain with > 4 heavy atoms)
                        then  (F)
                        else  if    (bulky side chain)
                              then (I)      else  (L)
else  if    (small)
      then  if    (side chain with < 4 heavy atoms)
            then  if    (linear side chain)
                  then  (T)    else  (S)
            else  if    (charged side chain)
                  then  (D)    else  (N)
      else  if    (side chain with O)
            then  if    (charged side chain)
                  then  (E)    else  (Q)
            else  (R)
(a)

If    (charged side chain)
then  if    (negativ charged)
      then  if    (small)
            then  (D)    else  (E)
      else  (R)
else  if    (small)
      then  if (hydrophobic)
            then  if    (side chain with < 3 CH)
                  then  (G)
                  else  if    (bulky side chain)
                        then  (V)    else  (A)
            else  if    (side chain with < 3 CH)
                  then  if    (side chain with < 4 heavy atoms)
                        then  (S)    else  (N)
                  else  (T)
      else  if    (hydrophobic)
            then  if    (side chain with > 4 heavy atoms)
                  then  (Y)
                  else  if    (bulky side chain)
                        then  (I)    else  (M)
            else  (Q).
(b)
```

Fig. 6. Interpretations of the decision trees from Fig. 5. (a) Fig. 5a. (b) Fig. 5b.

no histidine in position i+1. A detailed analysis of the instances of the two patterns (H_2 and H_6) of this type show that only 35 of the 76 and 53 instances, respectively, really occur at the beginning of α-helices in 24**

**The numbers in parentheses indicate the positions i: 156B (81), 1ABP (110), 1BP2 (39), 1FDH (3), 1HDS (145), 1HHO (165), 1MBS (3), 1PP2 (38, 161), 1RHD (183), 1RNS (25), 1SBT (242), 1TIM (196, 231, 444), 2APP (139), 2CCY (102, 230), 2CPP (28), 2CYP (149, 163, 231), 2LZM (59, 107), 2SNS (55, 62), 3C2C (49), 3CPV (59), 4APE (143), 4FDR (43, 96, 256), 4HHB (292, 435), 7CAT (468).

```
If    (side chain with < 3CH)
then  if    (side chain with < 4 heavy atoms)
      then  (S or G)      else  (D or N)
else  if    (side chain with < 4 heavy atoms)
      then  (T or P)
      else  if    (hydrophobic)
            then  if    (side chain with O)
                  then  (Y)    else  (L or K or H or R or F)
            else  (E)
```

Fig. 7. Interpretation of the decision tree from Fig. 4 ignoring occurrences of less than 10.

of the considered 84 proteins with respect to the assignment in [12]. The other instances describe kinks of long α-helices (H_2) and spacious turns (H_6), which are labeled by ...BTTB... in [12]. These observations are hints to look for further patterns, e.g. for kinks in long α-helices which are functionally important.

The interpretations of the decision trees from Fig. 5 are given in Fig. 6. As an example we consider the descriptions of the amino acids (A, G, V) following from the decision trees in Fig. 5. This set is described by the expression [(*hydrophobic*) AND (*small*) AND (*side chain without NH, OH, SH*)] (Fig. 5a) and by the expression [(*hydrophobic*) AND (*small*) AND [NOT (*charged side chaine*)]] (Fig. 5b). These expressions are the best descriptions with respect to the observed amino acid frequencies and the given properties. How these and other expressions are extracted from decision trees is obvious. Of course, this can be carried out automatically.

Ignoring occurrences of less than 10 the decision trees of Fig. 4 can be interpreted as shown in Fig. 7. It follows that the amino acid properties *side chain with < 3 CH*, *side chain with < 4 heavy atoms*, *hydrophobic*, and *side chain with O (oxygen)* are the most important properties for position i of type 1.

We have outlined a method which is a suitable tool for analyzing protein structure data. In particular, decision tree algorithms are able to uncover structurally induced amino acid partitions and to generate for that complex descriptions in which the logical operators AND, OR and NOT are used. The automatic incorporation of the operator NOT is of particular interest because negative constraints are important for structure description and deriving appropriate patterns for structure prediction.

REFERENCES

[1] Rooman, M.J. and Wodak, S. (1991) Proteins 9, 69–78.
[2] Taylor, W. (1986) J. Mol. Biol. 188, 233–258.
[3] Smith, R.F. and Smith, T.F. (1990) Proc. Natl. Acad. Sci. USA 87, 118–122.
[4] King, R.D. and Sternberg, J.E. (1990) J. Mol. Biol. 216, 441–457.
[5] Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) J. Mol. Biol. 112, 535–542.
[6] Unger, S. and Wysotzki, F. (1980) Lernfähige Klassifizierungssysteme, Akademie-Verlag, Berlin.

[7] Quinlan, R.J. (1984) in: Machine Learning – An Artificial Intelligenz Approach (Michalski, R.S., Carbonell, J.G. and Mitchell, T.M. eds.) pp. 463–482, Springer-Verlag, Berlin.
[8] Kaden, F., Koch, I. and Selbig, J. (1990) J. Theor. Biol. 47, 85–100.
[9] Haiech, J. and Sallantin, J. (1985) Biochimie 67, 555–560.
[10] Nakai, K., Kidera, A. and Kanehisa, M. (1988) Protein Engineering 2, 93–100.
[11] Richardson, J.S. and Richardson, D.C. (1988) Science 240, 1648–1652.
[12] Kabsch, W. and Sander, C. (1983) Biopolymers 22, 2577–2637.