New method for protein secondary structure assignment based on a simple topological descriptor

New Method for Protein Secondary Structure Assignment Based on a Simple Topological Descriptor

Todd Taylor, Margarita Rivera, Glenda Wilson, and Iosif I. Vaisman*
Laboratory for Structural Bioinformatics, School of Computational Sciences, George Mason University, Manassas, Virginia

ABSTRACT **A simple, five-element descriptor, derived from the Delaunay tessellation of a protein structure in a single point per residue representation, can be assigned to each residue in the protein. The descriptor characterizes main-chain topology and connectivity in the neighborhood of the residue and does not explicitly depend on putative hydrogen bonds or any geometric parameter, including bond length, angles, and areas. Rules based on this descriptor can be used for accurate, robust, and computationally efficient secondary structure assignment that correlates well with the existing methods. Proteins 2005;60:513–524. © 2005 Wiley-Liss, Inc.**

Key words: protein topology; secondary structure assignment; computational geometry; Delaunay tessellation

INTRODUCTION

Secondary Structure Assignment

Protein secondary structure assignment procedure is defined as a system of rules used to map local protein structure to one of several discrete classes,[1,2] such as α-helix, π-helix, β-strand, β-turn, etc. Secondary structure assignment is widely used in various areas of bioinformatics. It determines fold assignment[3] (e.g. all α, α/β, etc) and therefore is important in structural genomics. It plays an important role in constructing sequence[4] and structure[5] alignments. Accurate and consistent assignment of secondary structure defines the quality of a training set for secondary structure prediction algorithms, and it is necessary to develop such algorithms with maximum predictive power.[6]

One of the first computational methods for secondary structure assignment was developed in the late 1970s by Levitt and Greer.[7] Other early studies focused on characterization of β-turns[8] and secondary structure assignment based on simplified C_α-only representation of protein structures.[9] The DSSP method[10] introduced in 1983 became the *de facto* standard in secondary structure assignment. It assigns secondary structure to one of eight states mainly from putative main-chain hydrogen bonding patterns. Bonds are assumed to exist if the electrostatic energy between donor and acceptor groups satisfies the relation $E = kq_1q_2(1/r_{\mathrm{ON}} + 1/r_{\mathrm{CH}} + 1/r_{\mathrm{OH}} + 1/r_{\mathrm{CN}}) < -0.5$ kcal/mol, where q_1 and q_2 are partial charges. The eight states are defined as follows: α-helix (H), defined by a hydrogen bond between residues $i - 1$ and $i + 3$ as well as between i and $i + 4$; 3-10 helix (G), defined by a hydrogen bond between residues $i - 1$ and $i + 2$ as well as between i and $i + 3$; π-helix (I), defined by a hydrogen bond between residues $i - 1$ and $i + 4$ as well as between i and $i + 5$; bend (S), assigned to the central residue when the angle formed by residues $i - 2$, i, and $i + 2$ is greater than 70°; β-bridge (B), defined by hydrogen bonds between one pair of adjacent residues and another separated from the first; β-sheet (E), defined as two or more consecutive β-bridges; coil (C), defined as any residue to which none of the other class definitions apply. Often a reduced, three-letter secondary structure alphabet is used instead of the full eight-letter alphabet. One such reduced alphabet seen frequently is (H, G, I) → H; (B, E) → E; (C, S, T) → C; and this is the reduced alphabet we will use in this work. STRIDE[11] is a modified version of DSSP that assigns from the same eight letter alphabet as DSSP using putative hydrogen bonding patterns in addition to known preferred ranges of ϕ and ψ dihedral angles. As in DSSP, a putative hydrogen bond energy between donor (NH) and acceptor (CO) is defined and it has the form $E_{\mathrm{hb}} = E_r \times E_t \times E_p$, where E_r is a function of N–O distance, E_p is a function of the angle formed by N, H, and O, and E_t is a function of the angles defined by the hydrogen atom and bisectors of lone pair orbitals.[11] Helices and strands are designated when E_{hb} multiplied by a term reflecting the probability of finding the given torsion angles in the helix or strand regions of the Ramachandran plot exceeds a cutoff value. Secstr[12] is another modified version of DSSP, which uses a different definition of the π-helix. The first π-helical residue i has the CO group involved in a hydrogen bond with residue $i + 5$, and the last helical residue j has the NH group hydrogen bonded to residue $j - 5$, which also must be in the helix. Putative hydrogen bonds are defined in essentially the same way as with DSSP.

Several other secondary structure assignment methods are based on concepts other than DSSP. P-Curve[13] calcu-

Grant sponsor: National Science Foundation; Grant number: ITR/MCB 0112896 (a collaborative project with Alex Tropsha and Tamar Schlick).

M. Rivera's present address is Universidad Nacional Autónoma de México, Instituto de Química, Cd. Universitaria, 04510 México DF, México.

*Correspondence to: Iosif I. Vaisman, Laboratory for Structural Bioinformatics, School of Computational Sciences, George Mason University, 10900 University Blvd. MSN 5B3, Manassas, VA 20110. E-mail: ivaisman@gmu.edu

Received 20 September 2004; Revised 9 December 2004; Accepted 10 December 2004

Published online 10 May 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20471

lates a helicoidal axis from the atomic coordinates of the peptide backbone designed to have minimal curvature and to minimize the changes in the position of successive monomers with respect to the axis. The translations and rotations necessary to map a local coordinate system placed on each peptide bond to another local system placed on the helicoidal axis form a parameter set, and different secondary structure states have different characteristic ranges of these parameters. A circular dichroism (CD) spectroscopy related method, XTLsstr[14] relies on amide–amide backbone interaction, since far-UV protein CD spectra are determined largely by these interactions. It classifies residue conformations based on two angles and three distances: the angle ζ, defined by the carbonyl vectors C$\rightarrow$O of residues $i - 1$ and i; the angle τ, defined by α-carbons $i - 1$, i, and $i + 1$; dison3, the hydrogen-bonding distance measured from O(i) to N(i+3); dison4, the hydrogen-bonding distance measured from O(i) to N(i+4); discn3, the non-hydrogen-bonded distance from C(i) to N(i+3). There is no electrostatic definition of hydrogen bonding as with DSSP. XTLsstr assigns secondary structure to one of the seven states α-helix (H), 3-10 helix (G), β-strand (E), hydrogen-bonded β-turn (T), non-hydrogen-bonded turn (N), 3-1 helix (P), and coil (C).

The methods described above require coordinates of all peptide backbone atoms. There are several other methods that use only α-carbon coordinates. DEFINE[15] finds exact matches to ideal straight helices and strands using *linear distance masks*, template patterns in C_α–C_α distance matrices. Exact matches are then extended until the root mean square (RMS) difference between template and match in the query structure exceeds a cutoff. Bent helices and strands are then identified by pasting together straight helices and strands identified in the previous step if the angle between their axes is sufficiently small. P-SEA[16] also relies on α-carbon coordinates only and assigns conformations using the reduced alphabet H, E, and C. It uses distances $d2$, $d3$, and $d4$ between α-carbons $i - 1$, i, $i + 1$, and $i + 2$ as well as the angle τ defined by α-carbons $i - 1$, i, $i + 1$, and the dihedral angle α, defined by α-carbons $i - 1$, i, $i + 1$, and $i + 2$. A small number of rules map these three distances and two angles to secondary structure states.

In the VoTAP method,[17] a protein structure is first Voronoi tessellated, which results in an irregular polyhedron centered on each residue. The tessellation is performed using a relaxation process with three fictitious layers of molecules surrounding the protein, which ensures that surface polyhedra are closed. A residue contact matrix is then constructed. The matrix element is 2 if the irregular polyhedra corresponding to each residue are touching and share a large face; the element is 1 if the polyhedra are touching and share a small face, and the element is 0 if they do not touch. For every position i, contact motifs between i and residues $i - 6$ to $i - 2$, as well as $i + 2$ to $i + 6$, are extracted from the matrix. Mappings of these motifs to secondary structure have been constructed based on a consensus secondary structure assignment by DSSP, DEFINE, P-SEA, and STRIDE to a large non-redundant training set. Three state secondary structure assignments are made to novel structures by applying the mappings.

The Delaunay tessellation, used in the methods described in this work, is the *dual* of the Voronoi tessellation used by VoTAP,[18] meaning that there is one Delaunay simplex vertex for each Voronoi polyhedron and one Delaunay simplex edge for each Voronoi polyhedron face. The Voronoi tessellation of a point set can be derived from the Delaunay and vice versa. Our Delaunay-based assignment methods are therefore more closely related to VoTAP than to the other methods discussed here, but there are significant differences. VoTAP looks at characteristic patterns in two-body contacts between residue i and residues $i - 6$ to $i - 2$ and $i + 2$ to $i + 6$. Our methods look at patterns in four-body contacts with no restriction on separation in primary sequence. There are weights associated with contacts in VoTAP. With our methods, contacts are unweighted. VoTAP is constructed to simultaneously minimize conflicts with DSSP, STRIDE, P-SEA, and DEFINE. Our Delaunay-based assignment methods presented here are currently designed to match DSSP assignments only.

Comparison of the assignments obtained by different methods shows only fair agreement,[19] and work continues on improving the procedures, using both the traditional definition of secondary structure and new, unconventional definitions. For instance, DSSPcont[20] makes conventional eight-letter DSSP assignments for 10 values of the putative hydrogen bond energy cutoff, not just the usual -0.5 kcal/mol. A residue can have different assignments under different cutoffs. A 10-element vector $\mathbf{v}$ (corresponding to the ten cutoff levels) can be assigned to each residue for each secondary structure class S. It contains a 1 if the residue was assigned to that class under that cutoff and a 0 if it was not. The dot product of $\mathbf{v}$ and a vector of carefully chosen weights $\mathbf{w}$ gives a number that characterizes the degree to which this residue has secondary structure character consistent with S. A descriptor of eight numbers (corresponding to the eight DSSP states) can therefore be assigned to each residue, giving a kind of secondary structure character spectrum, rather than a single discrete state. The topological method described in this article can be also used to design a new secondary-structure assignment scheme.

Delaunay Tessellation of Protein Structures and Definition of *t*-Numbers

Protein structures can be analyzed using a computational geometry technique known as Delaunay tessellation. In this approach, each amino acid residue is abstracted to a point. This point can be collocated with either an α-carbon, a β-carbon, or the center of mass of a residue side chain. In the current work, we place the points at the α-carbon atoms. The structure is then tessellated in a unique way to form a set of non-overlapping, irregular, space-filling tetrahedra, whose vertices form a Delaunay simplex; the procedure by which they are generated is called Delaunay tessellation[18] (Fig. 1). The tetrahedra have the property that a sphere on the surface of which all four vertices reside does not contain a vertex from any

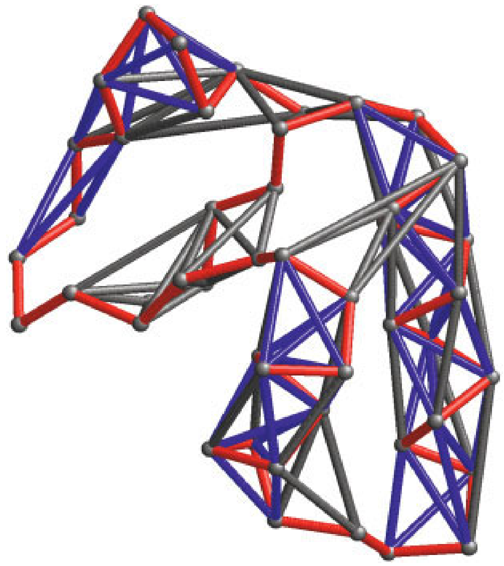

Fig. 1. Delaunay tessellation of crambin (1crn). Only simplices of types 2 and 4 are shown. Protein backbone is shown in red, simplices of type 2 in grey, type 4 in blue.

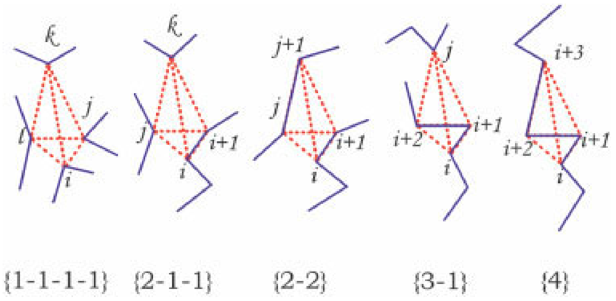

Fig. 2. Five types of Delaunay simplices classified by the way the main chain passes through the simplex. From left to right: type 0, in which no two vertex residues are consecutive in primary sequence; type 1, in which exactly one pair of vertex residues are consecutive; type 2, in which two pairs are consecutive but the pairs are separated from each other in primary sequence; type 3, in which exactly three vertex residues are consecutive in primary sequence; type 4, in which all four vertex residues are consecutive.

other tetrahedron (the *empty sphere property*). Residues joined by a simplex edge are considered nearest neighbors. Delaunay tessellation of protein structures has been used in fold recognition,[21–23] as the basis for a protein structure comparison algorithm,[24] as a way to identify cavities in the surface of a protein that could be potential binding pockets,[25] as a way to study stability effects of point mututations,[26,27] and as a basis to define structural motifs.[28–30]

A residue sequence proximity based classification of Delaunay simplices has been introduced[21] in which simplices resulting from the tessellation of protein structures were divided into five classes based on the way the main chain threads through them (Fig. 2). In type 0 simplices, none of the four residues at the vertices of the simplex are consecutive in primary sequence. In type 1 simplices, exactly one pair is consecutive in primary sequence. In type 2, two pairs are consecutive, but the pairs are separated from each other in primary sequence. In type 3, exactly three residues are consecutive. In type 4, all four residues are consecutive. One can tally the number of simplices of each type to which a given residue belongs, and we call these sums t-*numbers*. For example, if residue i is a vertex in 10 type 0 simplices, seven type 1 simplices, eight type 2 simplices, zero type 3 simplices, and four type 4 simplices, its t-numbers are $t0(i) = 10, t1(i) = 7, t2(i) = 8, t3(i) = 0$ and $t4(i) = 4$. The total number of simplices of all types in which a residue can participate varies greatly. In the analyzed dataset, this number ranges from 1 to 72.

MATERIALS AND METHODS

A set of 996 non-homologous X-ray structures with resolution ≤2.2 Å, crystallographic r-factor ≤0.23, and maximum pairwise sequence identity ≤30% was culled from the PDB using the PISCES web server.[31] This set will be referred to as *996culled*. The chains composing *996culled* were between 10 and 1231 residues long with a mean of 230. All of the structures were tessellated using the Qhull program[32] and software written by Zhibin Lu. The t-numbers for all residues in the data set were computed from the tessellations. The tessellation software requires coordinates for all C_α atoms without gaps, and no member of *996culled* had a missing C_α. Typically only about two thirds of PDB X-ray structures[33] can be Delaunay tessellated due to missing C_α coordinates.

Mappings from a feature vector consisting of t-numbers and their sums and differences to a secondary structure assignment were constructed. Some of these mappings were simply manually determined rules, while others were constructed using machine-learning software. One such machine-learning method was Quinlan's C4.5 decision tree package.[34] Each node in a decision tree specifies a test of one attribute in the feature vector **v**, and each branch descending from the node corresponds to one of the possible values of the attribute. A secondary structure class is associated with each leaf node. Objects are classified by starting at the root and at each node testing the attribute specified by the node and proceeding down the tree branch corresponding to the value of the attribute in **v**. Attributes that have higher *information gain* values (a quantity related to Shannon entropy) are tested at nodes higher in the tree, at or near the root. The information gain with respect to some classifying attribute A of the objects in collection of objects S is the expected reduction in entropy caused by partitioning S according to that attribute. High information gain attributes are the most distinguishing features and give the most power to classify. Another machine learning method used was RandomForest, the R package implementing Breiman's Random Forests.[35,36] With Random Forests, many different decision trees (a forest) are constructed using random subsets of the training data and random elements of the feature vector. To classify a new object, the feature vector is presented to all trees in the forest. Each tree gives a classification, and the overall classification chosen has the most votes over all the trees in the forest. The final machine learning mapping was built using *nnet*, the R implementation of a feed-forward back-propagation neural network.[35] In our case,

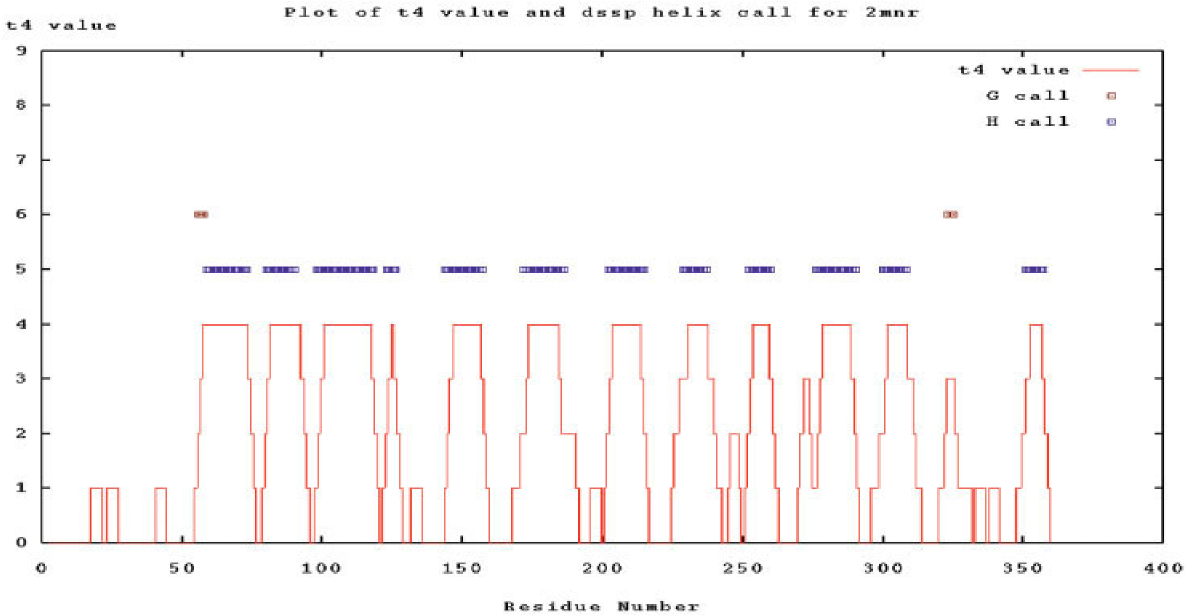

Fig. 3. Plot of *t*4 values along the length of PDB structure 2mnr with bars above indicating DSSP G assignment (upper) and DSSP H assignment (lower).

```
t4:     0000000000000001111001111000000000000011110000000000123444444444444444
DSSP:   CCEEEEEEEEEEEECCCCCEEECCEEECCEEEEEEEEEECCCCEEEEEEECCCHHHHHHHHHHHHHHHHH

t4:     4321001234444444444432100123444444444444444443210123432100011110000000
DSSP:   HCCCECCHHHHHHHHHHHHHHHECCEHHHHHHHHHHHHHHHHHHHHHCCCHHHHCCCCCECCEEEEECCC

t4:     0123444444444432100000000111223444444444443222221000011110123444444444
DSSP:   CCHHHHHHHHHHHHHCCCCEEEEEECCCHHHHHHHHHHHHHHHHHHCCEEEEEECCCCCHHHHHHHHHHH

t4:     4321000000001223334444444332110012221012344444432100000001233211234444
DSSP:   HHHCEEEEEEECECCCCHHHHHHHHHHHCEEECCCCCCCHHHHHHHHHHCEEEEEEEECCCCCCHHHHHH

t4:     4444443210000111223444444433211000000011233321111011110111100000011233
DSSP:   HHHHHHHCCEEEEEECCCCHHHHHHHHHCCCCEEEECCCCCCCCCCCCCCCCCCCCCCCEECCCEECCHH

t4:     4444321
DSSP:   HHHHHCC
```

Fig. 4. *t*4 Value and three-letter DSSP assignment for the PDB structure 2mnr.

the net had 15 input nodes, either nine or 15 hidden nodes, and one output node. The programs DSSP, DSSPcont, DEFINE, P-SEA, secstr, STRIDE, XTLsstr, and VoTAP were used to generate secondary structure assignments, which were compared to the results from Delaunay tessellation-based assignments. All calculations were performed on 195 MHz SGI Octane and 2.8G Hz Dell Precision computers. The list of 996 proteins in the set used in this paper and secondary structure assignments for all these proteins are located at http://binf.gmu.edu/struct_binf_group/SEC_STR/.

RESULTS AND DISCUSSION

Patterns in *t*-Numbers

After calculating the *t*-numbers for all of the residues in out dataset, we plotted *t*-numbers for each chain alongside known secondary structure assignments. These plots immediately reveal correlations and strong patterns (Figs. 3 and 4). For instance, the minimum possible value $t4(i)$ of the $t4$ number of residue i is 0 and the maximum value is 4. Recall that the definition of a type 4 simplex means that all four vertex residues must be consecutive in sequence. The only consecutive residue quadruplets which contain residue i (abbreviated r_i) are: $q1 = (r_{i-3}, r_2, r_{i-1}, r_i)$; $q2 = (r_{i-2}, r_{i-1}, r_i, r_{i+1})$; $q3 = (r_{i-1}, r_i, r_{i+1}, r_{i+2})$; $q4 = (r_i, r_{i+1}, r_{i+2}, r_{i+3})$. If none of these quadruplets forms a tetrahedron, then $t4(i) = 0$. If all four form tetrahedra, then $t4(i) = 4$. The intermediate cases are also possible. Hence $t4$ can only take on the values 0 to 4. As another example, it is always true that $|t4(i + 1) - t4(i)| \leq 1$ (cf. with the stair-step profiles in Fig. 3). To prove this, notice that given $t4(i)$, there are four possibilities for $t4(i + 1)$: (1) $q1$ forms a simplex and quadruplet $q5 = (r_{i+1}, r_{i+2}, r_{i+3}, r_{i+4})$ does not, in which case $t4(i + 1) = t4(i) - 1$; (2) $q1$ forms a simplex as does $q5$, in which case $t4(i + 1) = t4(i)$; (3) $q1$ does not form a simplex and neither does $q5$, in which case

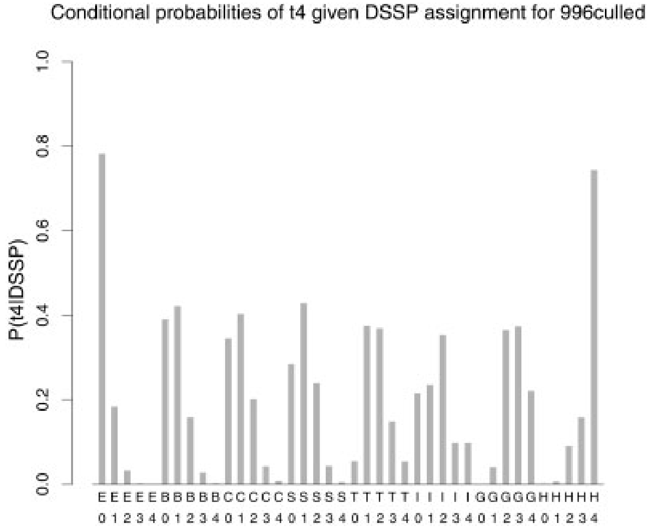

Fig. 5. Plot of P($t4$|DSSP) for eight-letter DSSP alphabet.

$t4(i + 1) = t4(i)$; (4) q1 does not form a simplex but $q5$ does, in which case $t4(i + 1) = t4(i) + 1$. Hence $t4$ always changes by at most 1 from one residue to the next.

t-Numbers and Secondary Structure

The patterns that can be observed in Figure 4 suggest that $t4 = 4$ is a very strong indicator of an α-helix and $t4 = 0$ a strong indicator of a β-strand. Furthermore, Figure 5 shows that residues in each DSSP category are distributed around an average '$t4$-ness,' and the distribution is unimodal. For example, the most probable $t4$ value of β-bridge residues (B) is 1, and the probability falls off monotonically away from $t4 = 1$. The situation is similar for $t2$ and $t3$, but the correlations are weaker. Helices have low $t3$ values, helices and strands tend to have high $t2$ values, while coils and turns have low $t2$ values. A reasonable interpretation of $t4$ might be a measure of compactness or helicity and $t2$ might be interpreted as a measure of local regularity of structure. These observations are confirmed by the data in Table I, which show correlations between t-numbers and the components of DSSPcont.[20] For example, there is a correlation of 0.81 between $t4$ number and the H component of DSSPcont which measures the 'degree' of α-helicity of a residue.

We conducted a series of experiments to determine whether an accurate t-number to DSSP mapping is possible. As a first approximation, we constructed by hand a system of context-free rules (the assignment to residue i only depends on the t-numbers of residue i) and applied them to *996culled* (there were too few π-helix (I) residues to generalize to sensible rules for this class). In the case of multiple assignments, the order of precedence was as follows: H > E > G > B > T > S > C. The rules, which individually give good sensitivity, appear in Table II(a). Table II(b) shows the accuracy of the mappings for a 7-letter DSSP alphabet as well as the reduced DSSP alphabet (G, H, I) → H, (B, E) → E, (C, S, T) → C. It can be noted that $t0$ and $t1$ are largely absent from the rules. They do not seem to be strong indicators of secondary structure. Geometries of the various simplex types may indicate the reasons for this. Unlike type 0 and type 1 simplices, the average edge length of types 2, 3, and 4 simplices is about 5-6 Å, typical values for C_α separation for residues joined by the hydrogen bonds that create secondary structure. Also, it should be noted that with the exception of H, E, and C, the agreement of context-free rules with DSSP is poor in the case of the full DSSP alphabet, but it improves considerably for the reduced, 3-letter alphabet.

Motif Based *t*-Number to DSSP Mappings

Figure 4 indicates that DSSP helices occur in unbroken stretches in $t4$ motifs, like [...123444...444321...] corresponding to [...HHHH...HHHH...], and that DSSP β-strands occur in $t4$ motifs, like [...21000...00012...] corresponding to [...EEEE...EEEE...]. Precisely where helices and strands terminate within these motifs often cannot be determined from $t4$ alone, especially for strands, but $t2$ and $t3$ values do give some indication of the boundaries. The rules we have settled on to call helices and strands within motifs follow. We refer to this method of assignment as *t4_motif*.

Helices:

(1) Find a stretch of at least three consecutive residues with $t4 = 4$, which will form a tentative helix core.
(2) Extend the helix boundary backward from the N-term core boundary until encountering a 'breaker' residue with either $t4 = 1$ or $t4 = 2$ and $t3 > 3$. The new N-term helix boundary is the residue immediately following the breaker in sequence.
(3) Extend the helix boundary forward from the C-term core boundary until encountering a 'breaker' residue with either $t4 = 1$ or $t4 = 2$ and $t3 > 3$. The new C-term helix boundary is the residue immediately preceding the breaker in sequence.
(4) All residues between the new helix boundaries are classified as helix.

Strands:

(1) Find a stretch of at least two consecutive residues with $t4 = 0$ bracketed by one or two residues with $t4 = 1$ on both sides. This stretch, including the flanking $t4 = 1$ residues, forms a tentative strand core.
(2) Push the strand boundary forward up to ⅓ of the length of the core from the N-term core boundary until encountering a 'starter' residue with $t2 + t3 \geq 10$. This is the new strand N-term boundary residue. If no such starter residue exists, then the N-term boundary remains unchanged.
(3) Push the strand boundary backward up to ⅓ of the length of the core from the C-term core boundary until encountering a 'starter' residue with $t2 + t3 \geq 8$. This is the new strand C-term boundary residue. If no such starter residue exists, then the C-term boundary remains unchanged.

TABLE I. Correlation of DSSPcont Scores and Solvent Accessible Area With *t*-Numbers for *996Culled*

	G	H	T	E	B	S	ACC
$t0$	−0.00187	−0.0111	−0.0125	0.108	−0.0305	−0.0429	−0.0898
$t1$	0.0785	0.137	0.0715	−0.273	−0.0194	0.0718	−0.0929
$t2$	−0.153	0.209	−0.248	0.383	0.0277	−0.181	−0.274
$t3$	−0.0947	−0.712	0.0193	0.513	0.0745	0.0961	−0.0797
$t4$	0.120	0.810	−0.0352	−0.565	−0.0882	−0.168	0.0835

TABLE II(A). Context Free Rules

B: $t0 \leq 9$ & $\mathrm{d}23 < 2$ & $5 \leq \mathrm{s}23 \leq 12$ & $\mathrm{d}24 \leq 6$ & $2 \leq \mathrm{s}24 \leq 8$ & $\mathrm{d}34 \leq 0$ & $4 \leq \mathrm{s}34 \leq 8$

C: $t4 = 0$ & $t2 \leq 3$

E: $t4 \leq 1$ & $|\mathrm{d}23| \leq 3$ & $\mathrm{s}23 \geq 6$

G: $t4 = 3$ & $t2 = 2$

H: $t4 = 4 \,\|\, (2 \leq t4 \leq 3$ & $t3 \leq 3$ & $\mathrm{d}23 \geq -1$ & $\mathrm{s}23 \leq 7)$

S: $(t4 = 0$ & $4 \leq t3 \leq 5$ & $2 \leq t2 \leq 3) \,\|\, (t4 = 1$ & $3 \leq t3 \leq 5$ & $1 \leq t2 \leq 3)$

T: $1 \leq t4 \leq 3$ & $\mathrm{d}23 \leq 0$ & $\mathrm{s}23 \leq 7$

Where: $\mathrm{s}23 = t2 + t3$; $\mathrm{d}23 = t2 - t3$; $\mathrm{s}24 = t2 + t4$; $\mathrm{d}24 = t2 - t4$; $\mathrm{s}34 = t3 + t4$; $\mathrm{d}34 = t3 - t4$.

(4) All residues between the new N-term and C-term for which $t0 > 1$ and $t2 > 2$ are classified as strand.

Coils:

(1) Every residue not classified as helix or strand is classified a coil.

Machine Learning Based *t*-Number to DSSP Mappings

Using manually determined rules, we could not reliably map *t*-numbers to the correct assignment for all eight structure categories in the standard DSSP alphabet. Does that mean that *t*-numbers do not contain enough information to construct this mapping, or that there are more subtle patterns? Are better mappings possible? In an attempt to answer these questions, we used the C4.5 decision tree package,[34] the R implementation of a three-layer feed forward back-propagation neural net with 15 hidden nodes,[35] and the R implementation of Breiman's Random Forests.[36] We trained each on *t*-number data from a randomly chosen subset of 20,000 residues taken from *996culled* with equal numbers of each of the secondary structure classes present. Again, there were too few π-helices assigned by DSSP to train on, so they were left out of the training set. We tested ≈224,000 residues from *996culled*. The resulting assignment methods will be referred to as *t*4_C4.5, *t*4_nnet, and *t*4_random_forest ('*t*4' because it is the strongest indicator of secondary structure among the *t*-numbers).

As with the manually determined rules under an eight-state alphabet, machine learning techniques could classify H, E, and C well, but not the other secondary structure classes [see Table III(a) e.g.]. Furthermore, it was found that training these methods with a feature vector consisting of not only $t0$–$t4$ but also the sums and differences $t2 \pm t3$, $t2 \pm t4$ and $t3 \pm t4$ as well as $t4$ of the previous two and following two residues improved results [compare Tables III(b and c) e.g.]. Except for Table III(b), all machine-learning methods presented here were trained on the latter 15-element feature vector.

The classification trees produced by C4.5 and Random Forests can be interpreted as a system of rules. Each node constitutes a binary split on one element of the feature vector; for instance, $t4 > 2$ or $t4 \leq 2$. The splits at the top of the tree near the root provide the most power to classify, while those near the leaves the least power.[37] Traversing the tree from root to leaf forms one classification rule consisting of a small set of these 'nodal' inequalities. The C4.5 trees we have built here are equivalent to between about 50 and 200 rules each consisting of between three and seven inequalities. Some examples of rules generated by the C4.5 decision trees and their accuracies are: $t0 > 10$ & $t1 \leq 7$ & $t2 = 2$ & $4 \leq t3 \leq 5$ → C (53.4%); $t2 > 3$ & $t4 > 3$ → H (88.6%); $t2 > 7$ & $t3 > 5$ → E (91.2%). For the trees trained on *t*-numbers only, all the splits up near the root of the tree involve $t4$. A bit lower down, they typically involve $t2$ and $t3$. Splits on $t1$ and $t0$ only come near the leaves. This confirms the observations from the manual rules discussed above, that $t4$ seems to provide the most information about secondary structure, followed by $t2$ and $t3$, with $t0$ and $t1$ contributing little information. Random Forest and C4.5 rule sets are significantly more complex than the manually determined rules, but they offer a more accurate *t*-number to DSSP mapping.

A comparison of secondary structure assignments according to the Delaunay tessellation-based methods we have described above, along with assignments from DSSP, DEFINE, P-SEA, secstr, STRIDE, VoTAP, and XTLsstr for three structures 1a12A, 1axn, and 2mnr, chosen as examples, is shown in Figure 6. Not surprisingly, the DSSP variants STRIDE and secstr are in closer agreement with DSSP, but Delaunay tessellation-based assignment agrees at about the same level as the other methods and is well within the range of agreement of different assignment methods reported elsewhere.[19] Tables IV(a–c) show agreement of assignment with DSSP as measured by Q3 and SOV.[38] Again, *t*-number-based methods are in the same range as existing methods. It is remarkable that a *t*-number-based secondary structure assignment method can be constructed that is consistent with existing methods, since the *t*-number descriptors reflect C_{α} backbone topology and connectivity. They have no explicit dependence on any lengths, angles, areas, or putative hydrogen bonds.

TABLE II(B). Results of Context-Free Rules for *996Culled*

	B	C	E	G	H	S	T
Sensitivity (7-letter)	0.134	0.224	0.791	0.010	0.940	0.000	0.094
Specificity (7-letter)	0.875	0.923	0.812	0.999	0.892	0.999	0.964
Sensitivity (3-letter)	—	0.534	0.783	—	0.893	—	—
Specificity (3-letter)	—	0.859	0.820	—	0.920	—	—

[sensitivity = TP/(TP + FN), specificity = TN/(TN + FP)]

TABLE III (A). Confusion Matrix for Seven-State Random Forest Based Assignment for *996Culled*

	B (DSSP)	C (DSSP)	E (DSSP)	G (DSSP)	H (DSSP)	S (DSSP)	T (DSSP)
B (r. forest)	1920	6422	7209	246	473	2601	1882
C (r. forest)	151	12558	3440	176	142	2634	1515
E (r. forest)	413	5701	31696	17	11	2127	649
G (r. forest)	58	2790	309	5910	5328	1323	4769
H (r. forest)	3	368	11	1529	64791	254	2278
S (r. forest)	192	8240	3731	283	382	6811	3678
T (r. forest)	140	5564	1669	1222	3339	3896	11864
Sensitivity	0.093	0.609	0.780	0.288	0.936	0.292	0.428
Specificity	0.995	0.856	0.910	0.983	0.937	0.936	0.924

TABLE III (B). Confusion Matrix for Three-State C4.5 Based Assignment Using only t0–t4 for *996Culled*

	C (DSSP)	E (DSSP)	H (DSSP)
C (C4.5)	60960	16404	10560
E (C4.5)	11011	39615	316
H (C4.5)	7770	973	75106
Sensitivity	0.693	0.778	0.896
Specificity	0.717	0.843	0.867

TABLE III (C). Confusion Matrix for Three-State C4.5 Based Assignment with 15-Element Feature Vector

	C (DSSP)	E (DSSP)	H (DSSP)
C (C4.5)	62505	7592	6773
E (C4.5)	19779	43252	221
H (C4.5)	7158	110	77368
Sensitivity	0.699	0.849	0.917
Specificity	0.894	0.885	0.948

TABLE III (D). Summary of Three-State Confusion Matrix Data for All Assignment Methods for *996Culled*

Assignment Method	C (DSSP)	E (DSSP)	H (DSSP)
*t*4_motif Sensitivity	0.670	0.771	0.908
*t*4_motif Specificity	0.859	0.894	0.918
*t*4_C4.5 Sensitivity	0.699	0.849	0.917
*t*4_C4.5 Specificity	0.894	0.885	0.948
*t*4_nnet Sensitivity	0.736	0.810	0.921
*t*4_nnet Specificity	0.882	0.905	0.947
*t*4_rand_forest Sensitivity	0.731	0.833	0.920
*t*4_rand_forest Specificity	0.890	0.900	0.950
DEFINE Sensitivity	0.354	0.928	0.915
DEFINE Specificity	0.933	0.734	0.902
P-SEA Sensitivity	0.778	0.773	0.853
P_SEA Specificity	0.823	0.909	0.969
secstr Sensitivity	0.934	0.821	0.999
secstr Specificity	0.931	0.999	0.958
STRIDE Sensitivity	0.920	0.988	0.978
STRIDE Specificity	0.982	0.987	0.965
XTLsstr Sensitivity	0.773	0.651	0.943
XTLsstr Specificity	0.836	0.941	0.922

Consistency of Assignment

It has been suggested that a good secondary structure assignment method ought to assign states in a very similar way to members of the ensemble of models in an NMR structure.[20] As a check of consistency and sensitivity to small changes in coordinates, we did secondary structure assignments for the first 10 models in each of the PDB NMR structures 1b2tA, 1cdn, 1d5vA, 1e41A, 1fsp, 1vreA, and 1xoa taken from Andersen et al.[20] We have calculated mean pairwise Q3 for the set of ten assignments for each structure for each assignment method (Table V). Q3, where the 3 refers to the three-letter secondary structure alphabet (C, E, H), is defined as the number of positions in which a pair of assignments agree divided by the length of the protein sequence and is identical to sensitivity as defined in Table II(b). We chose Q3 because it was used to test the consistency of DSSPcont.[20] Table V shows that *t*-number-based assignments are somewhat less consistent on these NMR structures than DSSP, STRIDE, and secstr but in the same range as DEFINE, P-SEA, and VoTAP, which, like the *t*-number methods, make assignments based on C_α coordinates only.

Dupuis et al. conducted a similar experiment comparing assignments for an obsolete, low-resolution structure and a newer, higher-resolution structure for the same protein.[17] They looked at 26 such structure pairs. The *t*-number methods require structures with no C_α gaps, the DSSP variants cannot work on C_α only structures, and DEFINE and VoTAP can miss assignment for some of the structures. Therefore, we analyzed a subset of 12 chains taken from these 26 for which we could get all methods to generate assignments without errors (151c and 351c; 1abk and 2abk; 1abp and 1abe; 1act and 2act; 1afnA and 2afnA;

```
 1 1a12A_dssp            100.0%  CCCCCCCCCCCCCEEEEEEEECCCCCCCCCCCCCEEEEEEEECCCCEEEEEEEECCCEEE
 2 1a12A_t4_motif         77.1%  CCCCCCCCCCCCEEEEEEEEEECCCCCCCCCCCCCEEEEEEEEECCEEEEEEEEEECCEEE
 3 1a12A_t4_c4.5          76.3%  CCCCCCHCCCCEEEEEEEEEEECCCCCCCCCHCCCCEEEEEECCEEEEEEEEEECCCEE
 4 1a12A_t4_nnet          75.1%  CCCCCCCCCCCEEEEEEEEEECCCCCCCCCHHCCCCEEEEEECCCEEECEEECCCCEE
 5 1a12A_t4_random_forest 77.6%  CCCCCCCCCCCEEEEEEEEEECCCCCCCCEHCCCCEEEEEECCCEEECCEEECCCEE
 6 1a12A_psea             75.3%  CCEEEEECCCCCEEEEEEEEEECCCCCCCCCCCCCCCEEEEEEECCCCEEEEEEECCEEE
 7 1a12A_secstr           84.5%  CCCCCCCCCCCCCEEEEEEEECCCCCCCCCCCCEECCCEECCCCCCCCCEEECCCEEE
 8 1a12A_stride           96.3%  CCCCCCCCCCCCCEEEEEEEECCCCCCCCCCCCEEEEEEEECCCCCEEEEEEECCCEEE
 9 1a12A_xtxtlsstr        73.3%  CCCCCCCCCCCCCEEEEEEECCCCCCCCCCCCCCCCCCCCCCCCCCCCCEEECCCEEE
10 1a12A_votap            71.3%  CCCCCCCCCCCCCCEEEECCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
   consensus/100%                CC......CCCC..EEEE....CCCCCCCCC...C.........C.........CC...

 1 1axn_dssp             100.0%  CCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHCCCCCCHHHHHHHHCCCCHHHHHHHHHHH
 2 1axn_t4_motif          92.6%  CCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHCCCCCCHHHHHHHHHHHCCCHHHHHHHHHH
 3 1axn_t4_c4.5           90.1%  CCCCCCCCECCCCCCCCCCHHHHHHHHHHHCCECCCCHHHHHHHHHCEEHHHHHHHHHHH
 4 1axn_t4_nnet           90.4%  CCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHCCCCCCCHHHHHHHHHHECHHHHHHHHHHH
 5 1axn_t4_rand_forest    92.6%  CCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHCCCCECCHHHHHHHHHHCCHHHHHHHHHHH
 6 1axn_define            84.2%  CCCCCCCCCEEEECCCCHHHHHHHHHHHHHHHCCCCHHHHHHHHHHHHHHHHHHHHHHHH
 7 1axn_psea              95.7%  CCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHCCCCCHHHHHHHHHCCCCHHHHHHHHHHH
 8 1axn_stride            98.8%  CCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHCCCCCHHHHHHHHHHCCHHHHHHHHHHH
 9 1axn_secstr            98.8%  CCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHCCCCCHHHHHHHHHHCCHHHHHHHHHHH
10 1axn_xtlsstr           89.8%  CCCCCCCCCEEECCCCCHHHHHHHHHHHHHHCCCCCHHHHHHHHHHCCHHHHHHHHHHHH
11 1axn_votap             95.4%  CCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHCCCCCCHHHHHHHHCCCCHHHHHHHHHHH
   consensus/100%                CCCCCCCCC.....CCCC.HHHHHHHHHHH...C.C.HHHHHHHH.....HHHHHHHHHH

 1 2mnr_dssp             100.0%  HHHHCCCCCCCEEEEEECCCCHHHHHHHHHHHHHHCCCCEEEEEECCCCHHHHHHHHHHH
 2 2mnr_t4_motif          76.8%  HHHHCCCCCECCEEEEEEECCCCHHHHHHHHHHHHHHCCCEEEEEEEECCHHHHHHHHHHH
 3 2mnr_t4_c4.5           78.7%  HHHHCCCECECCEEEEEEECECCHHHHHHHHHHHHHHECCEEEEEEECCCCHHHHHHHHHH
 4 2mnr_t4_nnet           81.2%  HHHHCCCEECCEEEEEEEEEECCHHHHHHHHHHHHHHCCCEEEEEEEECCCHHHHHHHHHH
 5 2mnr_t4_rand_forest    72.3%  HCCCCCCCCCCCHHCEECCCCCCHHHHHHHHHHCCHCCCCEEEECCCCCCHCHHHHHHH
 6 2mnr_define            69.2%  HHHHHEEEEEEEEEEEEEEEEECHHHHHHHHHHHHHHHHHHEEEEEEEEEEECCHHHHHHHHHHH
 7 2mnr_psea              81.2%  HHHHHCCEEEEEEEEEEEEECCCCHHHHHHHHHHHHHHHCCCEEEEEEECCCCHHHHHHHHHH
 8 2mnr_secstr            65.3%  CCHHHHCCCCCCCEEEEEEECCCCHHHHHHHHHHHHHHHCCCCEEEEEECCCCHHHHHHHHH
 9 2mnr_stride           96.4%  HHHHHCCCCCCEEEEEEECCCCCHHHHHHHHHHHHHHHCCCEEEEEECCCCHHHHHHHHHH
10 2mnr_xtlsstr           78.7%  HHHHHCCCCCCCCEEEECCCCHHHHHHHHHHHHHHHHCCCEEEEEECCCCHHHHHHHHHHH
11 2mnr_votap             79.0%  CCCCCCCEEEEEEEEEEEEECCCHHHHHHHHHHHHHHCCCCEEEEEECCCCHHHHHHHHHH
   consensus/100%                ..............EE...C..HHHHHHHHH..H......EE...CC....HHHHHHHH
```

Fig. 6. Portions of multiple sequence alignments of three-state secondary structure assignments to 1a12A (SCOP all-beta class), 1axn (SCOP all-alpha class), and 2mnr (SCOP alpha/beta class).

TABLE IV (A). Q3 and SOV With Respect to DSSP for 1a12A

	Q3				SOV			
	All	Helix	Strand	Coil	All	Helix	Strand	Coil
*t*4_motif	77.1	42.9	90.9	68.6	71.8	42.9	73.6	73.1
*t*4_C4.5	76.3	52.4	91.5	65.7	67.0	61.9	75.6	61.6
*t*4_nnet	75.1	66.7	84.7	67.6	63.6	71.4	63.3	63.2
*t*4_random forest	77.6	66.7	87.5	70.1	70.1	71.4	73.8	67.1
DEFINE	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
P-SEA	75.3	23.8	76.7	79.4	71.5	33.3	76.8	70.7
secstr	84.5	100.0	67.6	97.5	83.5	94.3	73.5	91.3
STRIDE	96.3	66.7	100.0	96.1	93.9	56.7	100.0	94.2
VoTAP	71.3	14.3	52.3	93.6	63.9	19.0	63.9	68.4
XTLsstr	73.3	85.7	55.1	87.7	68.2	82.9	65.2	69.3

1baa and 2baa; 4cnaA and 5cnaA; 1erl and 2erl; 2fnr and 1fnd; 2rhnA and 1aynA; 1sdhA and 3sdhA; 1trcA and 1fw4A). Q3 and SOV were computed for the structure pairs and then averaged over all 12 pairs for each assignment method. The results are compiled in Tables VI(a and b), which show that *t*-number methods are among the most consistent for overall assignment and for helix assignment and are comparable to the other methods for strands and coils.[19]

Locating Helix Caps

We have discussed the accuracy of helix boundary identification for a small set of three structures and the high-resolution–low-resolution pairs. It is well known that helix caps have a substantially different residue composition than proteins in general.[39] Therefore as a further test of the accuracy of helix boundary identification between tessellation-based assignment and DSSP, we computed the ratios $f_{i,\mathrm{N}}$ and $f_{i,\mathrm{C}}$, defined as the frequencies of amino acid i at the first and last positions of each helix of length seven or greater divided by the overall frequency of amino acid i for all of *996culled*. This gives a 20-element vector of frequency ratios for the N-terminal helix cap and another 20 element vector for the C-terminal cap, characterizing tendencies in cap amino acid composition. We computed these vectors for each assignment method and calculated

TABLE IV (B). Q3 and SOV With Respect to DSSP for 1axn

	Q3				SOV			
	All	Helix	Strand	Coil	All	Helix	Strand	Coil
*t*4_motif	92.6	96.5	N/A	83.0	97.9	100.0	N/A	92.6
*t*4_C4.5	90.1	92.6	N/A	84.0	84.8	91.9	N/A	71.1
*t*4_nnet	90.4	93.9	N/A	81.9	87.3	93.0	N/A	74.5
*t*4_random forest	92.6	96.9	N/A	81.9	83.9	89.4	N/A	73.3
DEFINE	84.2	100.0	N/A	45.7	75.3	85.8	N/A	54.0
P-SEA	95.7	96.1	N/A	94.7	94.6	93.2	N/A	98.6
secstr	98.8	99.6	N/A	96.8	99.7	100.0	N/A	98.9
STRIDE	98.8	99.6	N/A	96.8	95.2	93.4	N/A	100.0
VoTAP	95.4	94.3	N/A	97.9	94.3	92.6	N/A	98.9
XTLsstr	89.8	97.8	N/A	70.2	91.8	100.0	N/A	75.2

TABLE IV(C). Q3 and SOV with Respect to DSSP for 2mnr

	Q3				SOV			
	All	Helix	Strand	Coil	All	Helix	Strand	Coil
*t*4_motif	76.8	94.6	70.5	60.0	73.3	98.0	62.8	52.5
*t*4_C4.5	78.7	94.6	78.2	60.8	75.4	100.0	65.9	58.6
*t*4_nnet	81.2	96.0	83.3	63.1	78.3	100.0	67.8	63.9
*t*4_random forest	72.3	75.8	32.1	92.3	55.6	63.7	25.8	68.0
DEFINE	69.2	94.6	94.9	24.6	67.2	98.0	66.7	32.1
P-SEA	81.2	91.3	73.1	74.6	80.1	96.6	67.9	68.5
secstr	65.3	83.9	47.4	54.6	70.8	94.8	52.7	56.0
STRIDE	96.4	96.6	100.0	93.8	98.9	98.0	99.4	99.6
VoTAP	79.0	84.6	46.2	92.3	71.1	95.3	54.7	53.1
XTLsstr	78.7	96.0	65.4	66.9	76.6	98.0	61.8	63.4

TABLE V. Mean Pairwise Q3 for First Ten Models in Seven NMR Structures

	1b2tA	1cdn	1d5vA	1e41A	1fsp	1vreA	1xoa	mean
*t*4_motif	83.8	93.3	85.2	92.2	87.0	94.6	89.9	89.4
*t*4_C4.5	84.7	92.9	85.2	87.5	88.7	93.3	90.9	89.0
*t*4_nnet	85.8	91.6	85.8	86.0	87.0	93.4	91.5	88.7
DEFINE	89.0	86.9	88.8	91.3	92.2	92.2	90.6	90.1
DSSP	92.7	93.1	88.2	95.0	94.7	90.7	96.3	93.0
P-SEA	91.8	89.9	89.6	89.8	89.7	91.7	97.1	91.4
secstr	97.1	95.1	92.6	93.5	95.2	89.5	96.0	94.1
STRIDE	95.1	94.8	94.7	95.4	94.2	92.8	94.7	94.5
VoTAP	86.1	92.0	87.7	87.1	88.0	87.6	92.2	88.7
XTLsstr	83.1	85.2	81.9	87.7	80.6	84.4	88.5	84.5

TABLE VI (A). Mean Q3 of Twelve High Resolution/Low Resolution Pairs (Low wrt High)

	All	Helix	Strand	Coil
DEFINE	92.6	95.6	89.1	88.9
DSSP	89.9	86.4	78.9	95.4
P-SEA	91.2	92.6	76.9	95.0
secstr	91.9	86.8	91.4	96.7
STRIDE	90.6	86.8	83.6	97.1
*t*4_C4.5	91.8	95.2	86.0	90.5
*t*4_motif	92.1	94.7	72.4	92.8
VoTAP	90.6	89.7	90.6	93.1
XTLsstr	80.9	73.8	62.1	90.5

TABLE VI (B). Mean SOV of Twelve High Resolution/Low Resolution Pairs (Low wrt High)

	All	Helix	Strand	Coil
DEFINE	95.1	100.0	90.1	89.0
DSSP	87.2	91.2	71.7	86.1
P-SEA	89.4	95.6	78.8	86.4
secstr	90.4	90.8	92.8	89.7
STRIDE	88.6	90.6	78.8	87.4
*t*4_C4.5	84.1	95.9	83.7	78.7
*t*4_motif	90.1	98.7	74.4	86.3
VoTAP	88.5	92.5	90.9	88.5
XTLsstr	75.4	77.0	63.6	74.2

TABLE VII. Correlation of Helix Cap Frequency Ratios With Ratios of Rose and DSSP for *996Culled*

Method	N-Term (Rose)	C-Term (Rose)	N-Term (DSSP)	C-Term (DSSP)
DSSP	0.127	0.892	1.000	1.000
*t*4_motif	0.544	0.843	0.814	0.869
*t*4_C4.5	0.191	0.812	0.989	0.874
*t*4_rand_forest	0.181	0.868	0.990	0.910
DEFINE	0.772	0.343	0.076	0.196
P-SEA	0.089	0.895	0.987	0.928
secstr	0.137	0.897	0.995	0.904
STRIDE	0.183	0.874	0.988	0.963
XTLsstr	0.652	0.757	0.101	0.977

TABLE VIII. Helix and Strand Length Statistics for *996Culled*

	Mean Helix Length	SD Helix Length	Mean Strand Length	SD Strand Length	# Helices	# Strands
DSSP (3ltr)	9.416	6.282	4.385	3.010	8959	11729
DSSP (8ltr)	11.229	5.974	5.343	2.725	6688	9067
*t*4_motif	13.260	7.743	7.592	4.422	6637	7602
*t*4_C4.5	9.549	7.151	5.065	4.050	8863	12487
*t*4_rand_for	9.104	7.046	4.352	3.536	9300	13748
*t*4_nnet	9.588	7.021	4.478	3.459	8846	12879
DEFINE	14.562	8.554	8.478	9.386	3651	6614
P-SEA	11.978	5.800	6.325	2.357	6368	8742
secstr	10.736	7.076	4.442	2.208	8183	9518
STRIDE (3ltr)	10.311	6.872	4.522	3.082	8525	11729
STRIDE (8ltr)	12.532	6.283	5.427	2.807	6267	9262
XTLsstr	11.046	7.219	4.728	2.248	8239	9266

the linear correlation between them and (1) the DSSP vectors and (2) the data, where helix caps are defined by alternative hydrogen bonding patterns[39] (Table VII). For most of the methods, helix cap composition agrees well with DSSP. Not surprisingly, the putative H-bond-dependent methods, such as DSSP, STRIDE, and Secstr do not agree well with N-terminal helix cap compositions under the alternative H-bonding motifs,[39] some of which even involve hydrogen bonding to side-chain groups. Only DEFINE agrees well with data from ref. 39 at the N-term, with XTLsstr and the *t*4_motif method agreeing to a fair degree.

Secondary Structure Element Lengths and Helix Torsion Angles

The mean helix and strand lengths in *996culled*, according to several existing assignment methods as well as the *t*-number methods described earlier, are shown in Table VIII. It is interesting to note that with the C_α-only methods *t*4_motif, DEFINE, and P-SEA, helices and strands tend to be longer and fewer in number. However, this is not true in the machine-learning based *t*-number methods. The fact that they have shorter and more numerous secondary structure elements can be attributed to more complex mappings and exhaustive training to mimic DSSP.

Contour histograms/Ramachandran plots of ϕ-ψ angles for all of the residues assigned as helices by eight of the methods considered here are shown in Figure 7. In each case, the vast majority of the approximately 70,000 residues in each plot fall in the traditional right hand helix portion of the plot. However, with *t*4_motif, DEFINE, and XTLsstr, there are appreciable numbers of outliers with ψ greater than 100°. Table IX shows that the variance in ϕ and ψ angles drops dramatically as *t*4 increases. Residues with $t4 = 4$, usually found in the cores of helices, are restricted to the ‘bulls eye’ region of the plots in Figure 7. Of the $t4 = 4$ residues assigned helix by DSSP, STRIDE also assigns helix 99.8% of the time, but this rate of agreement drops to 80.8% when $t4 = 1$.

CONCLUSIONS

We have shown that reliable secondary structure information can be extracted from the Delaunay tessellation of a protein even though it is a minimal description of protein structure. Only C_α coordinates are required to derive it and it contains no explicit information about lengths, areas, angles, or any other arbitrary parameters. The *t*-number assignment methods are comparable, as measured by several criteria, with most other methods for secondary structure assignment. New, robust, topological definitions of helices and strands can be developed based on the rules described in this article. A more detailed and consistent description of secondary structure in terms of Delaunay tessellation-derived descriptors may be possible by, for instance, further subdividing the five classes of simplex described here and introducing new *t*-numbers, by excluding from the *t*-number sums simplices with long edges or large volumes, by mapping *t*-numbers under several values of edge length cutoff to secondary structure, or by including additional contextual information on the consecutive residues.

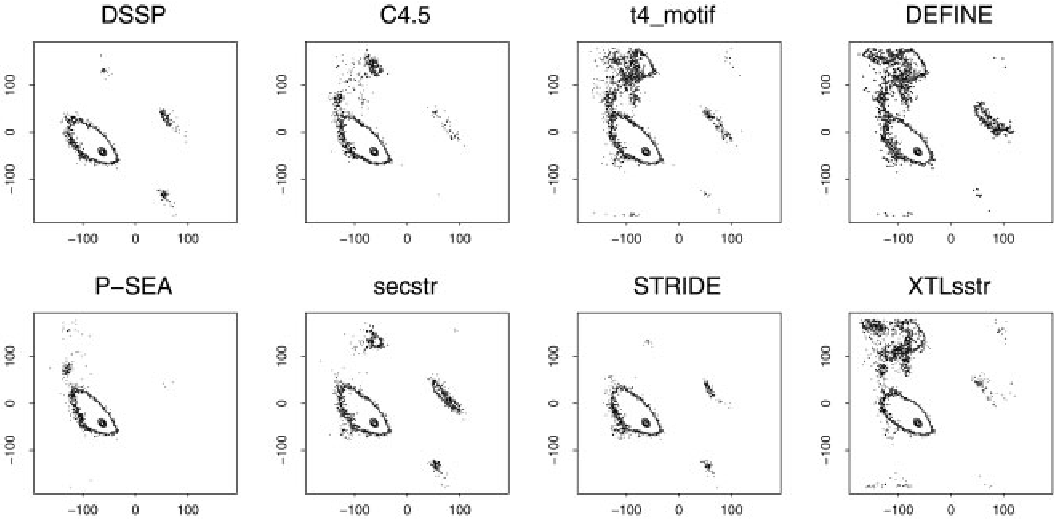

Fig. 7. Contour histogram/Ramachandran plot of ϕ-ψ values for residues assigned H by eight assignment methods. The bin size of the histograms is 2° square.

TABLE IX. Phi and Psi of DSSP H Residues in *996Culled* Broken Down by t4

*t*4	Size of Set	Fraction H w/STRIDE	Mean ϕ	SD ϕ	Mean ψ	SD ψ
0	101	0.861	−92.3	21.9	−12.8	20.7
1	433	0.808	−64.0	65.9	−33.9	67.3
2	6408	0.947	−69.2	23.0	−33.4	17.3
3	11448	0.970	−66.2	14.1	−35.6	12.6
4	55137	0.998	−64.3	8.0	−40.8	8.4

ACKNOWLEDGMENTS

This work was supported in part by a grant from the NSF. We thank John Grefenstette for helpful discussions and Zhibin Lu for several of the programs used in this work.

REFERENCES

1. Efimov AV. Standard structures in proteins. Prog Biophys Mol Biol 1993;60(3):201–239.
2. Andersen CA, Rost B. Secondary structure assignment. Methods Biochem Anal 2003;44:341–363.
3. Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res 2000;28(1):257–259.
4. Errami M, Geourjon C, Deleage G. Detection of unrelated proteins in sequences multiple alignments by using predicted secondary structures. Bioinformatics 2003;19(4):506–512.
5. Dror O, Benyamini H, Nussinov R, Wolfson H. MASS: multiple structural alignment by secondary structures. Bioinformatics 2003;19 Suppl 1:I95–I104.
6. Andersen CA, Bohr H, Brunak S. Protein secondary structure: category assignment and predictability. FEBS Lett 2001;507(1):6–10.
7. Levitt M, Greer J. Automatic identification of secondary structure in globular proteins. J Mol Biol 1977;114(2):181–239.
8. Isogai Y, Nemethy G, Rackovsky S, Leach SJ, Scheraga HA. Characterization of multiple bends in proteins. Biopolymers 1980; 19(6):1183–1210.
9. Ramakrishnan C, Soman KV. Identification of secondary structures in globular proteins—a new algorithm. Int J Pept Protein Res 1982;20(3):218–237.
10. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577–2637.
11. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins 1995;23(4):566–579.
12. Fodje MN, Al-Karadaghi S. Occurrence, conformational features and amino acid propensities for the pi-helix. Protein Eng 2002; 15(5):353–358.
13. Sklenar H, Etchebest C, Lavery R. Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. Proteins 1989;6(1):46–60.
14. King SM, Johnson WC. Assigning secondary structure from protein coordinate data. Proteins 1999;35(3):313–320.
15. Richards FM, Kundrot CE. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. Proteins 1988;3(2):71–84.
16. Labesse G, Colloc'h N, Pothier J, Mornon JP. P-SEA: a new efficient assignment of secondary structure from C alpha trace of proteins. Comput Appl Biosci 1997;13(3):291–295.
17. Dupuis F, Sadoc JF, Mornon JP. Protein secondary structure assignment through Voronoi tessellation. Proteins 2004;55(3):519–528.
18. Okabe A. Spatial tessellations : concepts and applications of Voronoi diagrams. Wiley series in probability and statistics 2000; xii:671.
19. Colloc'h N, Etchebest C, Thoreau E, Henrissat B, Mornon JP. Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. Protein Eng 1993;6(4):377–382.
20. Andersen CA, Palmer AG, Brunak S, Rost B. Continuum secondary structure captures protein flexibility. Structure (Camb) 2002; 10(2):175–184.
21. Singh RK, Tropsha A, Vaisman, II. Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. J Comput Biol 1996;3(2):213–221.
22. Tropsha A, Singh RK, Vaisman, II, Zheng W. Statistical geometry analysis of proteins: implications for inverted structure prediction. Pac Symp Biocomput 1996;614–623.
23. Krishnamoorthy B, Tropsha A. Development of a four-body statis-

tical pseudo-potential to discriminate native from non-native protein conformations. Bioinformatics 2003;19(12):1540–1548.

24. Bostick D, Vaisman, II. A new topological method to measure protein structure similarity. Biochem Biophys Res Commun 2003; 304(2):320–325.
25. Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S. Analytical shape computation of macromolecules: I. Molecular area and volume through alpha shape. Proteins 1998;33(1):1–17.
26. Masso M, Vaisman, II. Comprehensive mutagenesis of HIV-1 protease: a computational geometry approach. Biochem Biophys Res Commun 2003;305(2):322–326.
27. Carter CW Jr, LeFebvre BC, Cammer SA, Tropsha A, Edgell MH. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. J Mol Biol 2001;311(4):625–638.
28. Tropsha A, Carter CW Jr, Cammer S, Vaisman, II. Simplicial neighborhood analysis of protein packing (SNAPP): a computational geometry approach to studying proteins. Methods Enzymol 2003;374:509–544.
29. Cammer SA, Carty RP, Tropsha A. Computational methods for macromolecules: challenges and applications. Proc of the 3rd Intl Workshop on Algorithms for Macromolecular Modeling 2000;477–494.
30. Wako H, Yamato T. Novel method to detect a motif of local structures in different protein conformations. Protein Eng 1998; 11(11):981–990.
31. Wang G, Dunbrack RL, Jr. PISCES: a protein sequence culling server. Bioinformatics 2003;19(12):1589–1591.
32. Barber CB, Dobkin DP, Huhdanpaa H. The quickhull algorithm for convex hulls. ACM Transactions on Mathematical Software 1996;22(4):469–483.
33. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C. The Protein Data Bank. Acta Crystallogr D Biol Crystallogr 2002;58(Pt 6 No 1):899–907.
34. Quinlan JR. C4.5 : programs for machine learning. The Morgan Kaufmann series in machine learning 1993:x, 302 p.
35. Venables WN, Ripley BD. Modern applied statistics with S-plus. 1999.
36. Breiman L. Random forests—random features, Technical Report 567. 1999.
37. Mitchell TM. Machine learning. McGraw-Hill series in computer science Artificial intelligence 1997:xvii, 414 p.
38. Zemla A, Venclovas C, Fidelis K, Rost B. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins 1999;34(2):220–223.
39. Aurora R, Rose GD. Helix capping. Protein Sci 1998;7(1):21–38.