Structural alphabets as tools for the analysis of protein structures

Pr Bernard OFFMANN

Unité Fonctionnalité et Ingénierie des Protéines

CNRS FRE 3478 - Université de Nantes

Synopsis

✓Introduction

- Classical backbone description
- Structural alphabets
- A structural alphabet : Protein Blocks

✓ Structure analysis using structural alphabets

✓ Mining protein structures

✓ Analysis of structural diversity of pentapeptides in protein structures

✓Fold recognition

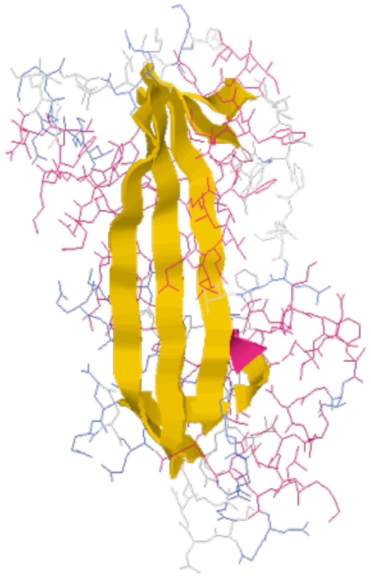
Classical description of backbone

Helices (28-35%)

↓

α helix
3.$_{10}$–helix
π helix

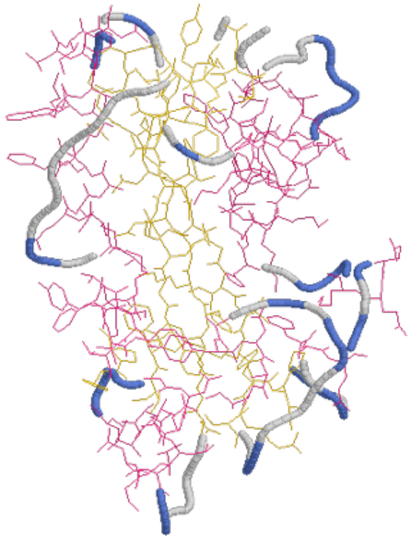
Polyproline II



Sheets (18-26%)

↓

β sheets
β strands
E strands
β bulges



Turns/coils (40-50%)

↓

γ, β, α, π turns
Ω loops
β hairpins
αα corners

Secondary structure assignment

Different assignment methods :

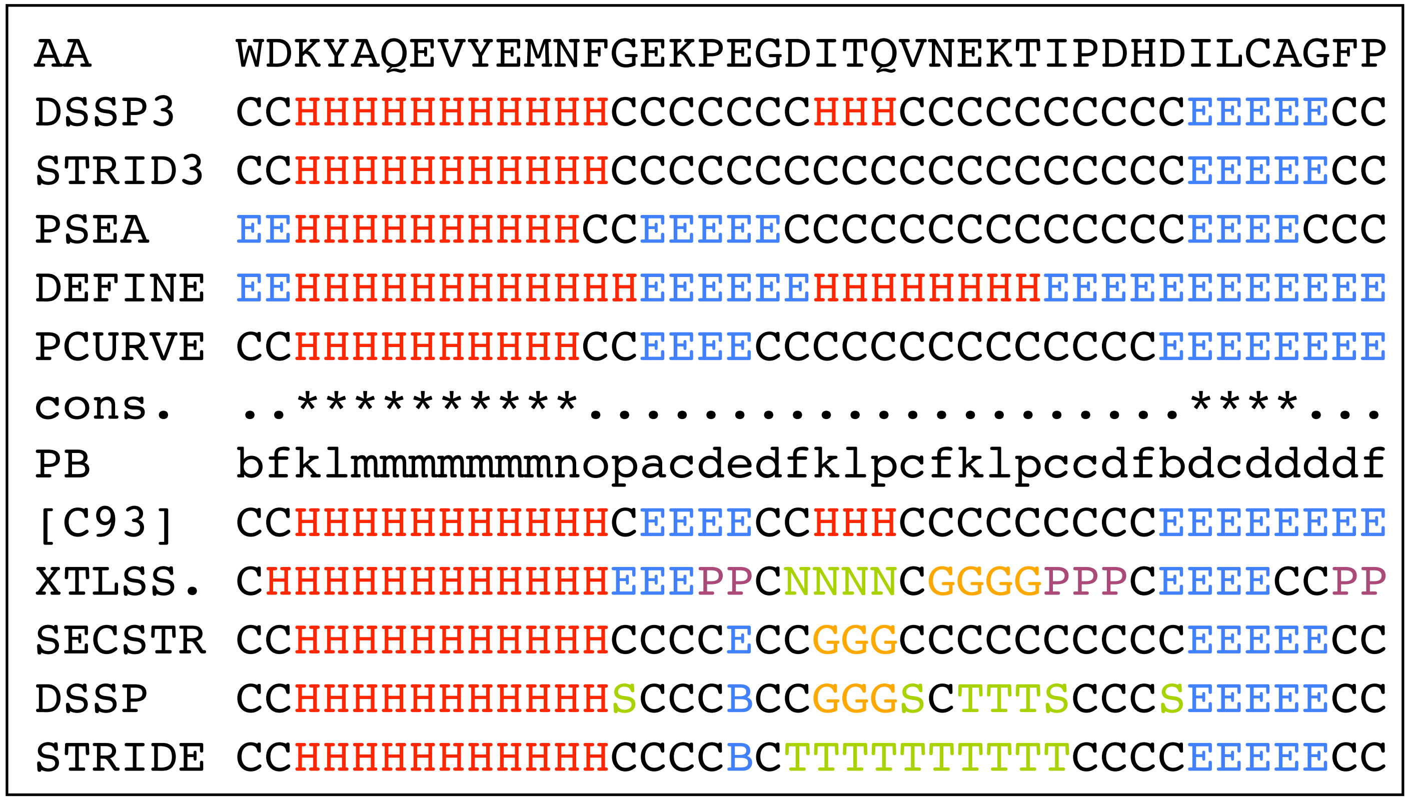
DSSP (Kabsch & Sander, 1983). H-bond

Secondary structure assignment

Different assignment methods :

Greer & Levitt (1977)	Distance
DSSP (Kabsch & Sander, 1983).	H-bond
DEFINE (Kundrot & Ridchards, 1988).	Distance
PCURVE (Sklenar, Etchebest and Lavery, 1989).	Axis
CONCENSUS (Colloc'h, Etchebest *et al*., 1993).	Mean
STRIDE (Frishmann & Argos, 1995).	H-bond / dihedral
PSEA (Labesse *et al*., 1997).	Distance / angle
PROSS (Srinivasan & Rose, 1999).	Dihedral
XTLSSTR (King & Johnson , 1999).	Distance / angle
DSSPcont (Andersen *et al*., 2001).	H-bond / dihedral
SECSTR (Fodje & Al-Karadaghi, 2002).	H-bond / dihedral
VORO3D (Dupuis *et al*., 2004).	Volume
KAKSI (Martin *et al*., 2005).	Distance / dihedral
SEGNO (Cubellis *et al*., 2005).	angle / multiple
Beta-Spider (2005), **PALSSE** (2005), **Delaunay tessalation** (2005)	

Secondary structure assignment

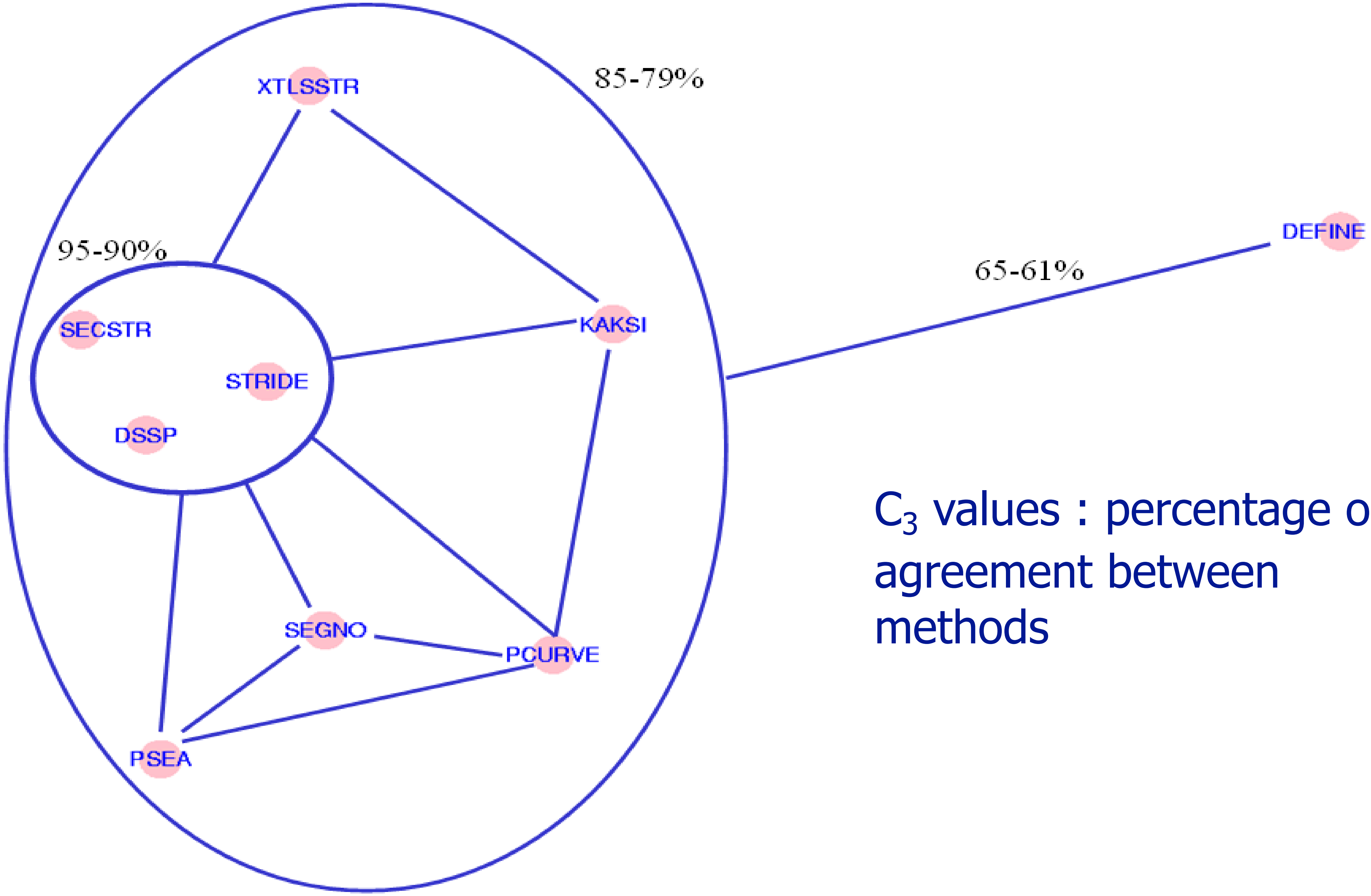


```
AA     WDKYAQEVYEMNFGEKPEGDITQVNEKTIPDHDILCAGFP
DSSP3  CCHHHHHHHHHHHCCCCCCCHHHCCCCCCCCCCEEEEECC
STRID3 CCHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCEEEEECC
PSEA   EEHHHHHHHHHHCCEEEEECCCCCCCCCCCCCCEEEECCC
DEFINE EEHHHHHHHHHHHHEEEEEEHHHHHHHHEEEEEEEEEEEE
PCURVE CCHHHHHHHHHHCCEEEECCCCCCCCCCCCCCEEEEEEEE
cons.  ..**********.....................****...
PB     bfklmmmmmmmnopacdedfklpcfklpccdfbdcddddf
[C93]  CCHHHHHHHHHHHCEEEECCHHHCCCCCCCCCEEEEEEEE
XTLSS. CHHHHHHHHHHHHEEEPPCNNNNCGGGGPPPCEEEECCPP
SECSTR CCHHHHHHHHHHHCCCCECCGGGCCCCCCCCCCEEEEECC
DSSP   CCHHHHHHHHHHHSCCCBCCGGGSCTTTSCCCSEEEEECC
STRIDE CCHHHHHHHHHHHCCCCBCTTTTTTTTTTCCCCEEEEECC
```

Example of secondary structure assignments for the protein 1OMH with DSSP, STRIDE, PSEA, DEFINE, PCURVE, XTLSSTR and SECSTR.

Fourrier, Benros & de Brevern (2004) *BMC Bioinformatics*, **5**, 58.
Offmann, Tyagi & de Brevern (2007) *Current Bioinformatics*, 2(3):1-38

Secondary structure assignment

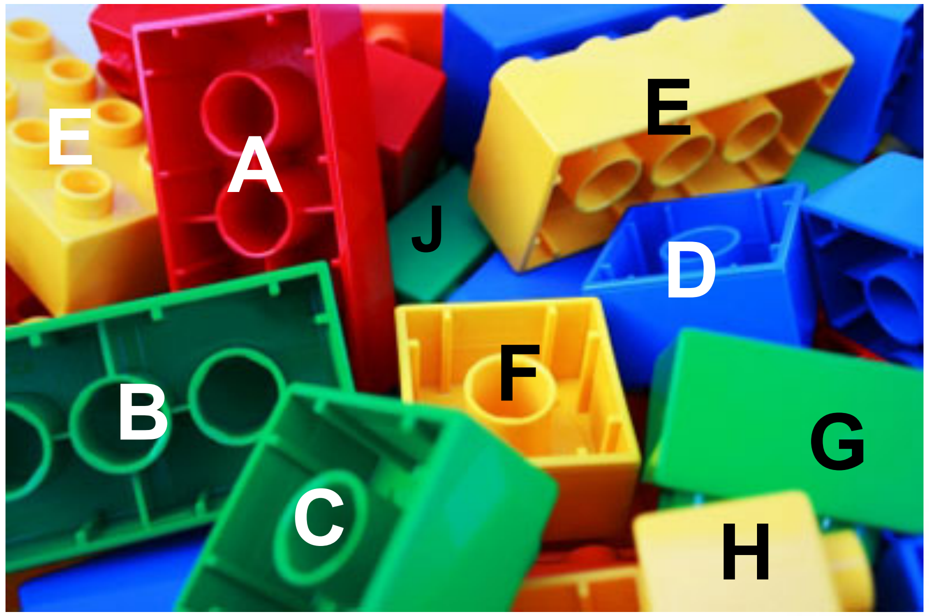


C_3 values : percentage of agreement between methods

Offmann, Tyagi & de Brevern, (2007). *Current Bioinformatics*, 2(3):1-38

UNIVERSITÉ DE NANTES

LEGO® bricks



Structural alphabets

A structural alphabet is a set (or library) of small prototypes which approximate every part of the protein structures.

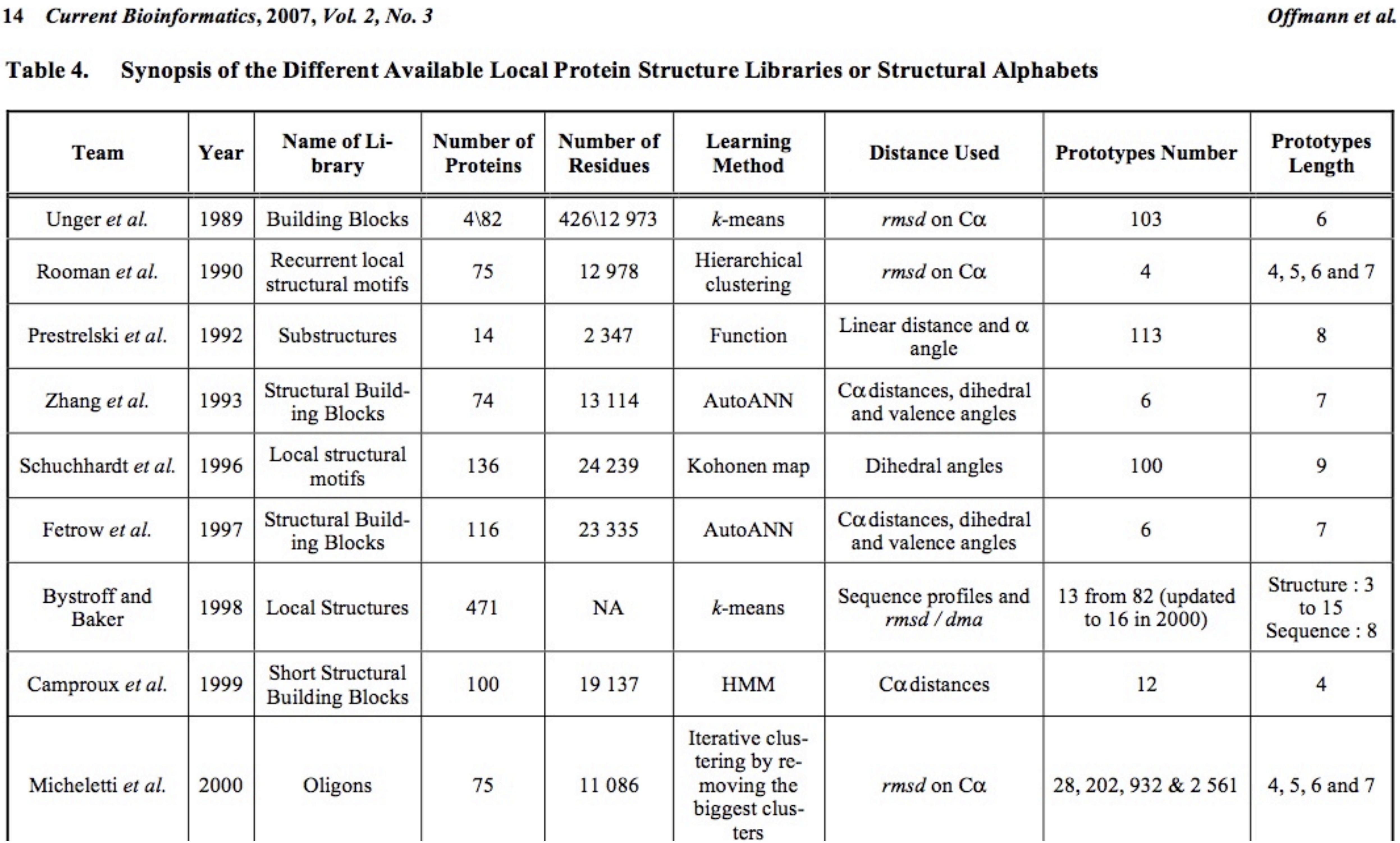
They are composed by a limited number of recurrent structural elements of proteins.

The associations between these structural "letters" are governed by logical rules and form the words of protein structures.

A structural alphabet has no a priori in regards to the secondary structures, i.e. it is not a categorization of the coil state.

Offmann, Tyagi & de Brevern, (2007). *Current Bioinformatics*, 2(3):1-38

Structural alphabets

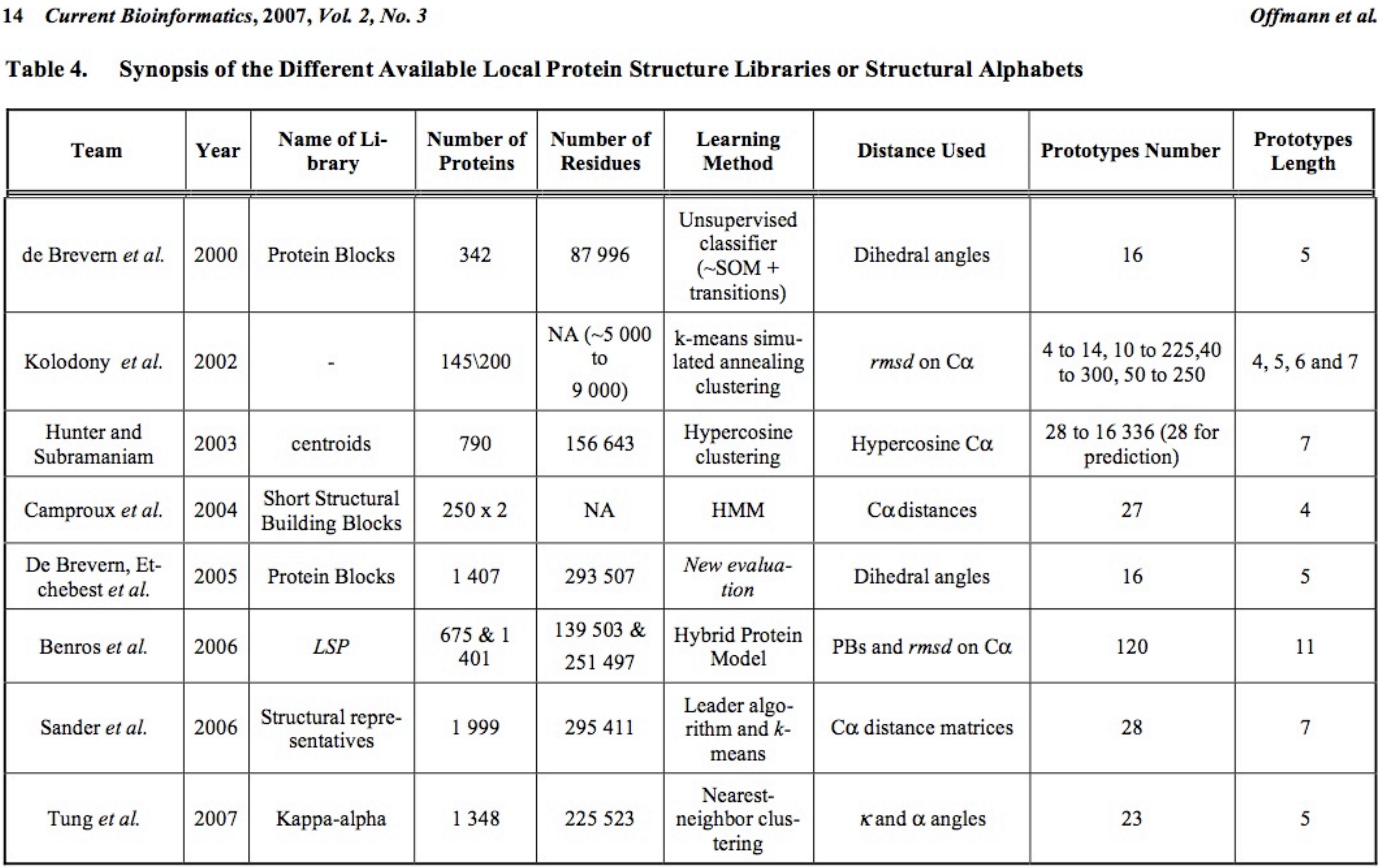


14 *Current Bioinformatics*, 2007, *Vol. 2, No. 3* *Offmann et al.*

Table 4. Synopsis of the Different Available Local Protein Structure Libraries or Structural Alphabets

Team	Year	Name of Library	Number of Proteins	Number of Residues	Learning Method	Distance Used	Prototypes Number	Prototypes Length
Unger *et al.*	1989	Building Blocks	4\82	426\12 973	*k*-means	*rmsd* on Cα	103	6
Rooman *et al.*	1990	Recurrent local structural motifs	75	12 978	Hierarchical clustering	*rmsd* on Cα	4	4, 5, 6 and 7
Prestrelski *et al.*	1992	Substructures	14	2 347	Function	Linear distance and α angle	113	8
Zhang *et al.*	1993	Structural Building Blocks	74	13 114	AutoANN	Cα distances, dihedral and valence angles	6	7
Schuchhardt *et al.*	1996	Local structural motifs	136	24 239	Kohonen map	Dihedral angles	100	9
Fetrow *et al.*	1997	Structural Building Blocks	116	23 335	AutoANN	Cα distances, dihedral and valence angles	6	7
Bystroff and Baker	1998	Local Structures	471	NA	*k*-means	Sequence profiles and *rmsd / dma*	13 from 82 (updated to 16 in 2000)	Structure : 3 to 15 Sequence : 8
Camproux *et al.*	1999	Short Structural Building Blocks	100	19 137	HMM	Cα distances	12	4
Micheletti *et al.*	2000	Oligons	75	11 086	Iterative clustering by removing the biggest clusters	*rmsd* on Cα	28, 202, 932 & 2 561	4, 5, 6 and 7

Structural alphabets

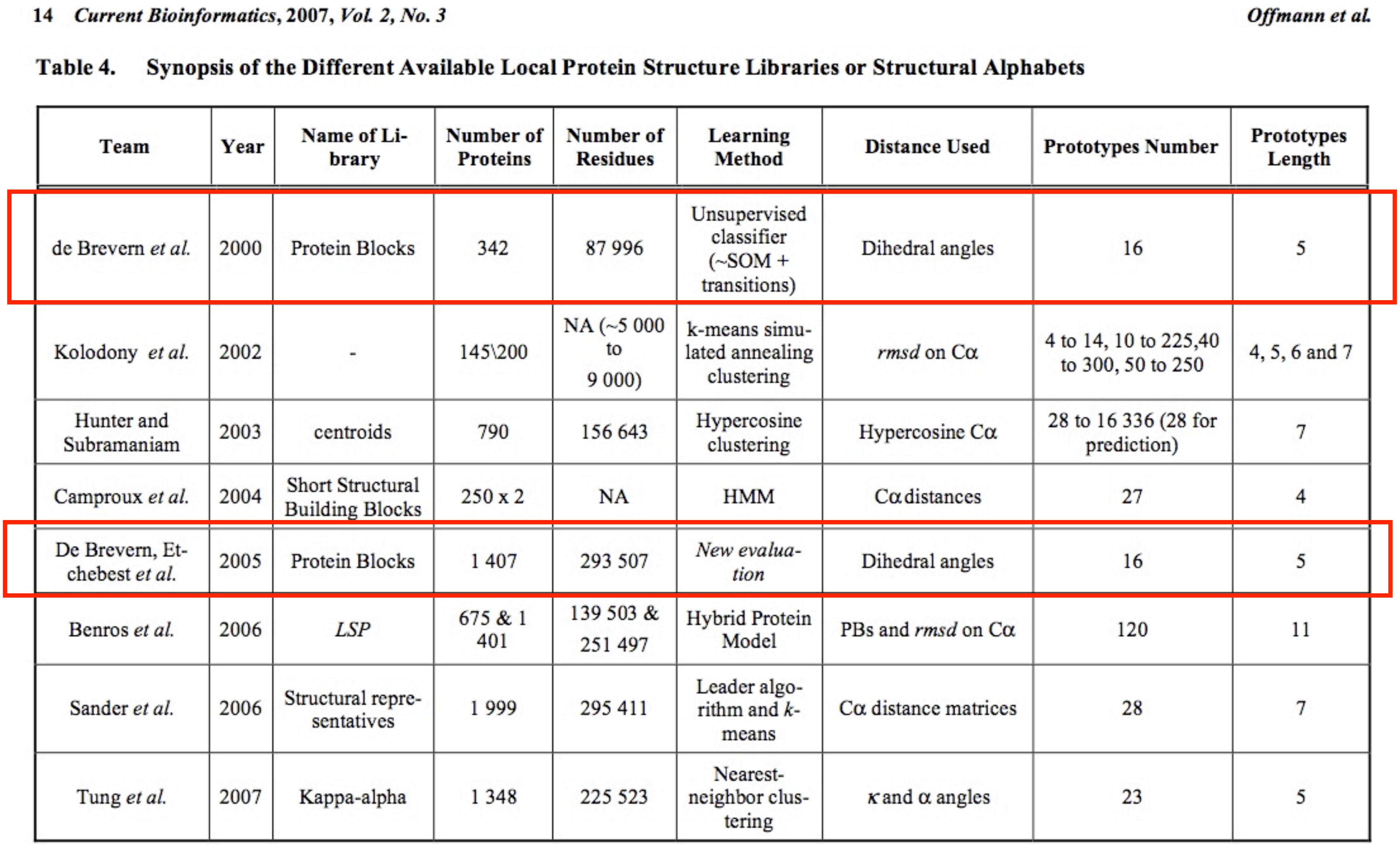


14 *Current Bioinformatics*, 2007, *Vol. 2, No. 3* *Offmann et al.*

Table 4. Synopsis of the Different Available Local Protein Structure Libraries or Structural Alphabets

Team	Year	Name of Library	Number of Proteins	Number of Residues	Learning Method	Distance Used	Prototypes Number	Prototypes Length
de Brevern *et al.*	2000	Protein Blocks	342	87 996	Unsupervised classifier (~SOM + transitions)	Dihedral angles	16	5
Kolodony *et al.*	2002	-	145\200	NA (~5 000 to 9 000)	k-means simulated annealing clustering	*rmsd* on Cα	4 to 14, 10 to 225,40 to 300, 50 to 250	4, 5, 6 and 7
Hunter and Subramaniam	2003	centroids	790	156 643	Hypercosine clustering	Hypercosine Cα	28 to 16 336 (28 for prediction)	7
Camproux *et al.*	2004	Short Structural Building Blocks	250 x 2	NA	HMM	Cα distances	27	4
De Brevern, Etchebest *et al.*	2005	Protein Blocks	1 407	293 507	*New evaluation*	Dihedral angles	16	5
Benros *et al.*	2006	*LSP*	675 & 1 401	139 503 & 251 497	Hybrid Protein Model	PBs and *rmsd* on Cα	120	11
Sander *et al.*	2006	Structural representatives	1 999	295 411	Leader algorithm and *k*-means	Cα distance matrices	28	7
Tung *et al.*	2007	Kappa-alpha	1 348	225 523	Nearest-neighbor clustering	κ and α angles	23	5

Structural alphabets



14 *Current Bioinformatics*, 2007, *Vol. 2, No. 3* *Offmann et al.*

Table 4. Synopsis of the Different Available Local Protein Structure Libraries or Structural Alphabets

Team	Year	Name of Library	Number of Proteins	Number of Residues	Learning Method	Distance Used	Prototypes Number	Prototypes Length
de Brevern *et al.*	2000	Protein Blocks	342	87 996	Unsupervised classifier (~SOM + transitions)	Dihedral angles	16	5
Kolodony *et al.*	2002	-	145\200	NA (~5 000 to 9 000)	k-means simulated annealing clustering	*rmsd* on Cα	4 to 14, 10 to 225,40 to 300, 50 to 250	4, 5, 6 and 7
Hunter and Subramaniam	2003	centroids	790	156 643	Hypercosine clustering	Hypercosine Cα	28 to 16 336 (28 for prediction)	7
Camproux *et al.*	2004	Short Structural Building Blocks	250 x 2	NA	HMM	Cα distances	27	4
De Brevern, Etchebest *et al.*	2005	Protein Blocks	1 407	293 507	*New evaluation*	Dihedral angles	16	5
Benros *et al.*	2006	*LSP*	675 & 1 401	139 503 & 251 497	Hybrid Protein Model	PBs and *rmsd* on Cα	120	11
Sander *et al.*	2006	Structural representatives	1 999	295 411	Leader algorithm and *k*-means	Cα distance matrices	28	7
Tung *et al.*	2007	Kappa-alpha	1 348	225 523	Nearest-neighbor clustering	κ and α angles	23	5

Structural alphabets

✓Prototypes libraries are useful for predicting protein backbone from sequence

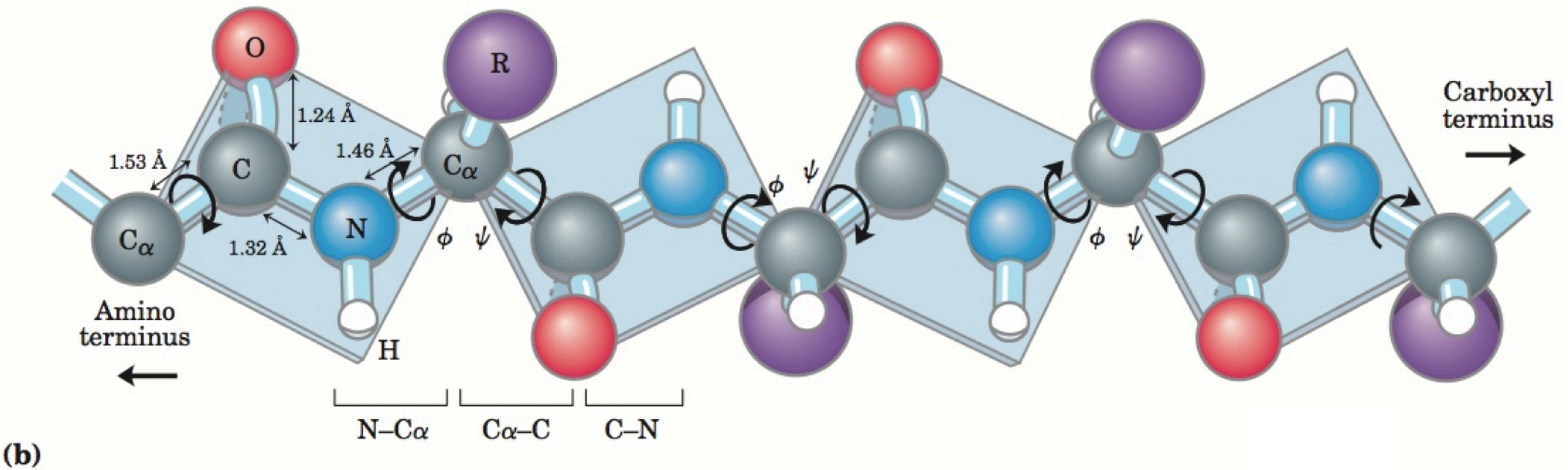
- e.g. I-Sites (Bystroff & Baker, 1996)

✓Number of states needs to be optimized both for precision of backbone description and prediction efficiency

✓Protein Blocks (de Brevern et al, 2000) has been developed towards this goal

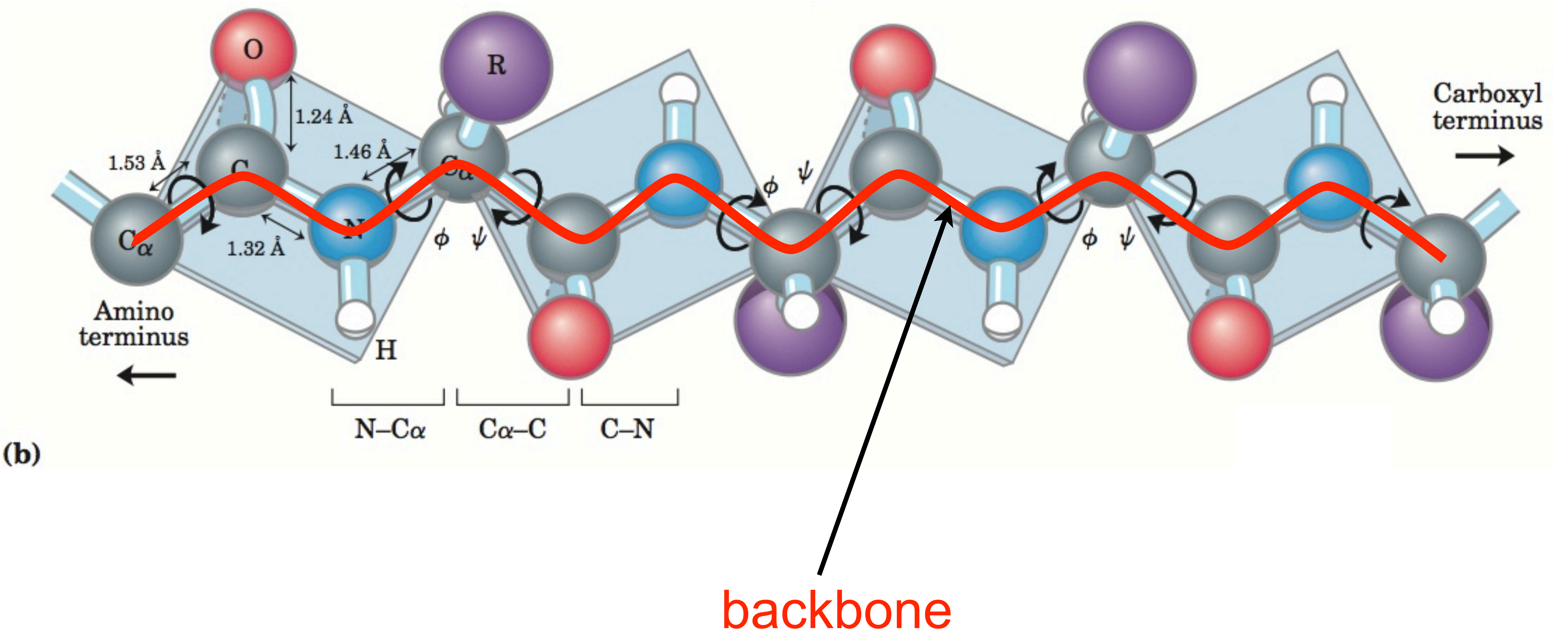
Protein Blocks

Set of 16 structural prototypes of 5 consecutive residues defined by specific phi and psi values



Protein Blocks

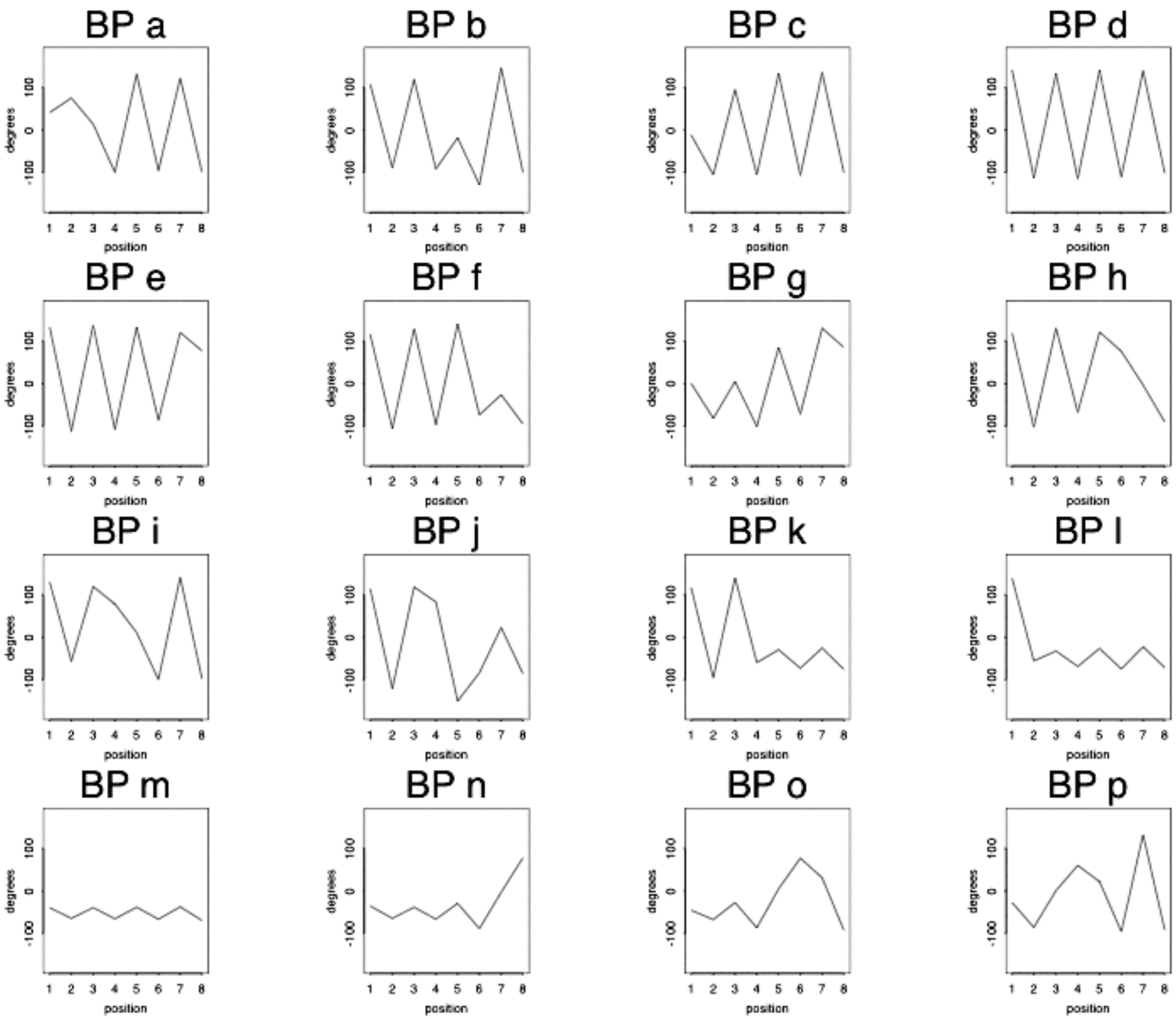
Set of 16 structural prototypes of 5 consecutive residues defined by specific phi and psi values



Protein Blocks

Set of 16 structural prototypes of 5 consecutive residues defined by specific phi and psi values

These were obtained by unsupervised classification of dihedral angles derived from unrelated protein structures using a self organizing map (SOM) also called Kohonen map.

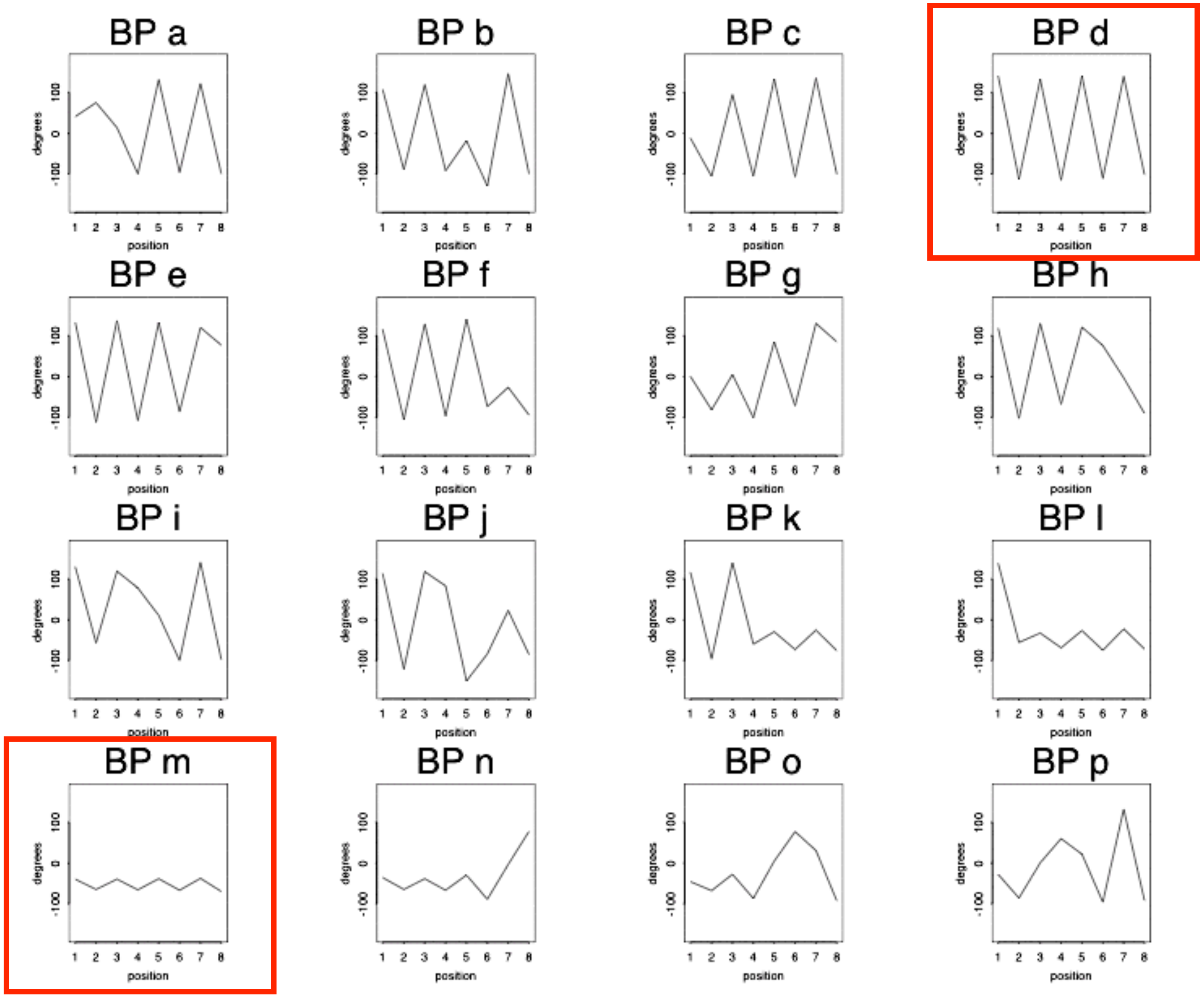


Courtesy : de Brevern

Protein Blocks

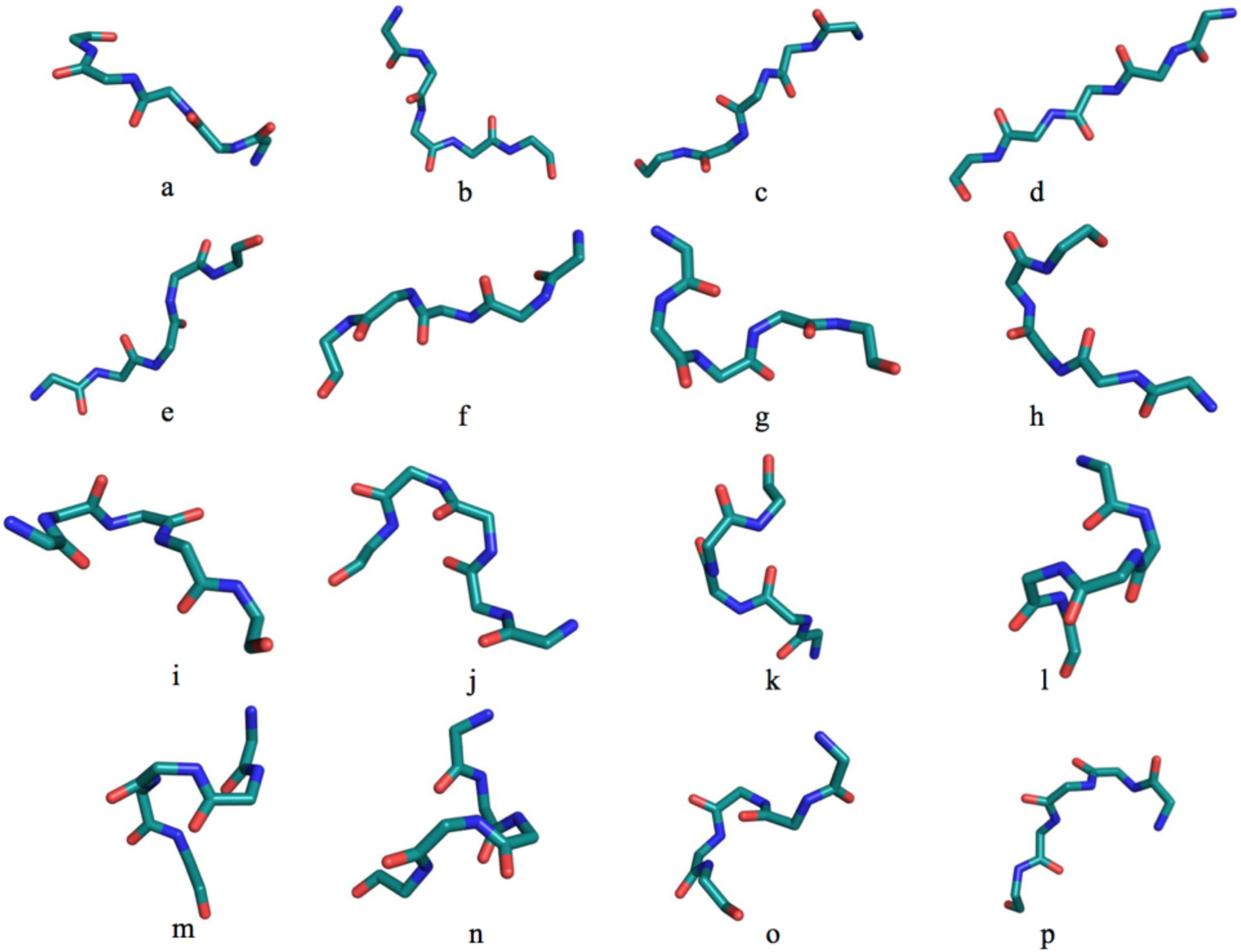
Set of 16 structural prototypes of 5 consecutive residues
defined by specific phi and psi values

These were obtained by unsupervised classification of dihedral angles derived from unrelated protein structures using a self organizing map (SOM) also called Kohonen map.



Courtesy : de Brevern

Protein Blocks


a
b
c
d
e
f
g
h
i
j
k
l
m
n
o
p