Practical course 3
Secondary structure prediction – conformational diseases

1. Analysis of secondary structure prediction programs
The programs that will be used to predict the secondary structure are GOR4, HNN and Sopma. They are provided on the website "http://npsa-pbil.ibcp.fr/" ("secondary structure prediction" section). Describe briefly the approach used by each program.

GOR IV (http://npsa-pbil.ibcp.fr/cgibin/npsa_automat.pl?page=npsa_gor4.html):
The GOR (Garnier, Osguthorpe, and Robson) method is an information-based method for the prediction of secondary structures of proteins. There is no defined constant decision.
GOR IV uses all possible pair frequencies within the window of 17 amino acid residues (Garnier et al.,1996). It was developed in the late 1970's shortly after the simpler Chou-Fasman method (Chou & Fasman, 1974). Like Chou-Fasman method, GOR method is also 92 Specific Secondary Structure Prediction Tools based on probability parameters derived from empirical studies of known protein tertiary structures solved by X-ray crystallography. The program gives two outputs: one is eye-friendly and gives the sequence and the predicted secondary structure in parallel rows, with symbols H= α helix, E=extended or β strand and C=coil; the second gives the probability values for each secondary structure at each amino acid position. The predicted secondary structure is the one with the highest probability-compatible structure with a predicted helix segment of at least four residues and a predicted extended segment of at least two residues.

HNN (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_nn.html):
The HNN (Hierarchical Neural Network) prediction method employs two networks to predict structures: a sequence-to-structure network and a structure-to-structure network. Thus, the prediction is only based on local information. Neural network methods are trained to recognize amino acid patterns by providing a data set containing known structures. The algorithm then identifies the structures present in unknowns
link to four videos which explain all running operations: https://www.youtube.com/playlist?list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi

Sopma (self-optimizedprediction from multiple alignment)

Recently a new method called the self-optimized prediction method (SOPM) has been described to improve the success rate in the prediction of the secondary structure of proteins. In this paper we report improvements brought about by predicting all the sequences of a set of aligned proteins belonging to the same family. This improved SOPM method (SOPMA) correctly predicts 69.5% of amino acids for a three-state description of the secondary structure (alpha-helix, beta-sheet and coil) in a whole database containing 126 chains of non-homologous (less than 25% identity) proteins. Joint prediction with SOPMA and a neural networks method (PHD) correctly predicts 82.2% of residues for 74% of co-predicted amino acids. Predictions are available by Email to deleage@ibcp.fr or on a Web page (http://www.ibcp.fr/predict.html).

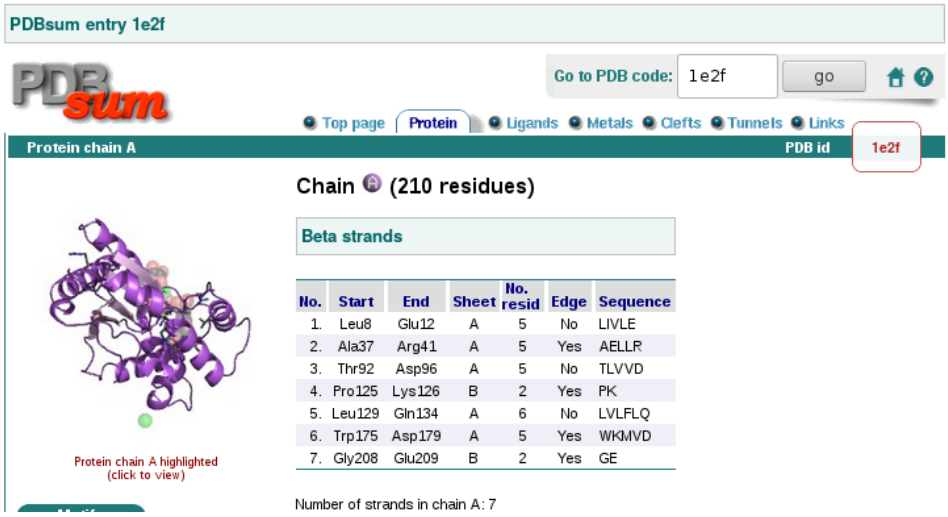
2. Comparison of the performances of secondary structure prediction programs

A] Search for the human thymidylate kinase (PDB code 1E2F) in the protein databank (www.rcsb.org). Search for the secondary structure limits of the protein in PDBSum (http://www.ebi.ac.uk/pdbsum; "Protein" tab, "7 strands" and "11 helices" in the menu on the left). This secondary structure assignation is provided in the file 1e2f.xls (on the Virtual University). A secondary structure assignation is also provided in the PDB file, sections "HELIX" and "SHEET" (from the PDB website, menu "Display file", "PDB file", find the "HELIX" and "SHEET" sections). Compare both secondary structure assignations (from PDBSum and from the PDB file). Why are they slightly different?

Result from PDB file:

```
HELIX    1   1 GLY A   18  ALA A   33  1                                   16
HELIX    2   2 THR A   47  GLN A   58  1                                   12
HELIX    3   3 GLU A   64  GLU A   78  1                                   15
HELIX    4   4 GLN A   79  GLN A   89  1                                   11
HELIX    5   5 TYR A   98  ALA A  108  1                                   11
HELIX    6   6 SER A  113  GLN A  119  1                                    7
HELIX    7   7 PRO A  120  VAL A  122  5                                    3
HELIX    8   8 GLN A  136  ALA A  141  1                                    6
HELIX    9   9 ASN A  153  MET A  168  1                                   16
HELIX   10  10 SER A  183  THR A  201  1                                   19
HELIX   11  11 ALA A  202  LYS A  205  5                                    4
SHEET    1   A 5 TRP A 175  ASP A 179  0
SHEET    2   A 5 LEU A 129  GLN A 134  1  N  VAL A 130   O  LYS A 176
SHEET    3   A 5 LEU A   8  GLU A  12  1  N  VAL A  10   O  LEU A 129
SHEET    4   A 5 THR A  92  ASP A  96  1  N  LEU A  93   O  ILE A   9
SHEET    5   A 5 ALA A  37  ARG A  41  1  N  GLU A  38   O  THR A  92
```

result from PDBSum:



No.	Start	End	Sheet	No. resid	Edge	Sequence
1.	Leu8	Glu12	A	5	No	LIVLE
2.	Ala37	Arg41	A	5	Yes	AELLR
3.	Thr92	Asp96	A	5	No	TLVVD
4.	Pro125	Lys126	B	2	Yes	PK
5.	Leu129	Gln134	A	6	No	LVLFLQ
6.	Trp175	Asp179	A	5	Yes	WKMVD
7.	Gly208	Glu209	B	2	Yes	GE

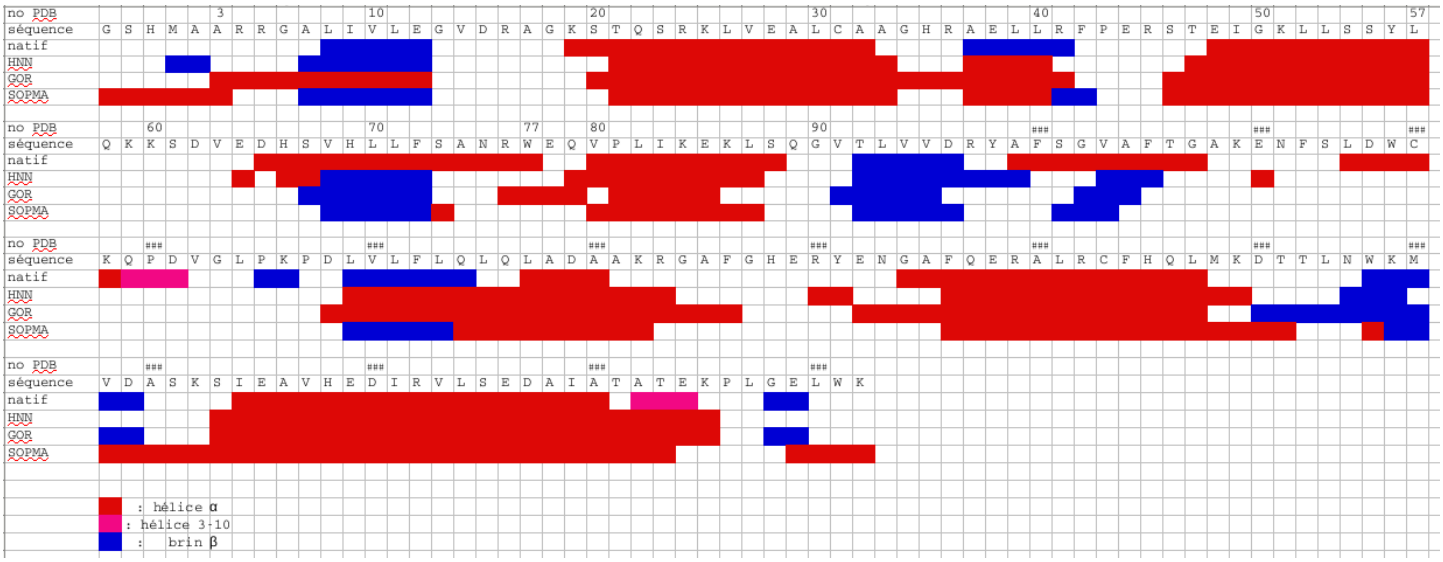
Table of helices

No.	Start	End	Type	No. resid	Length	Unit rise	Residues per turn	Pitch	Deviation from ideal	Sequence
*1.	Lys19	Ala32	H	14	21.10	1.49	3.62	5.39	7.0	KSTQSRKLVEALCA
*2.	Glu48	Leu57	H	10	15.53	1.53	3.54	5.41	4.2	EIGKLLSSYL
*3.	Asp65	Trp77	H	13	19.92	1.52	3.50	5.31	8.4	DHSVHLLFSANRW
*4.	Val80	Ser88	H	9	13.88	1.49	3.67	5.46	8.4	VPLIKEKLS
5.	Ala99	Gly107	H	9	13.71	1.46	3.68	5.37	10.1	AFSGVAFTG
6.	Leu114	Lys118	H	5	8.12	1.50	3.62	5.44	21.7	LDWCK
7.	Gln119	Asp121	G	3	-	-	-	-	-	QPD
8.	Leu137	Ala140	H	4	7.10	0.71	7.79	5.57	83.7	LADA
9.	Gly154	Leu167	H	14	21.47	1.51	3.55	5.35	8.3	GAFQERALRCFHQL
10.	Ile184	Ala200	H	17	26.15	1.50	3.74	5.62	11.7	IEAVHEDIRVLSEDAIA
11.	Ala202	Glu204	G	3	-	-	-	-	-	ATE

Number of helices in chain A: 11

Concerning the beta strands, 2 of them lack in PDB file, those from the B sheet. For the helices of the PDB file, they are shorter of two amino acids, one before and one after. The PDB files are generated from DSSP (http://swift.cmbi.ru.nl/gv/dssp/) which looks at the hydrogen bonds for every four residues to check the presence of helices at a local level. PDBSum uses PROCHECK (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC29784/) to generate its results. In any case, the differences are present especially as regards the beginning and the end of these domains.

B] Perform a secondary structure prediction of 1E2F by using the three programs GOR4, HNN and Sopma. Show in the .xls file the results. Analyse and comment these results.



Hierarchical Neural Network result for : 1E2Fxx0

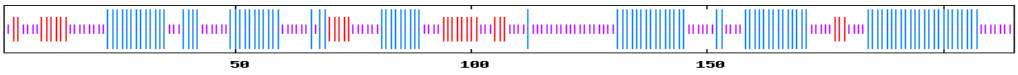
Abstract Guermeur, Y. PhD Thesis

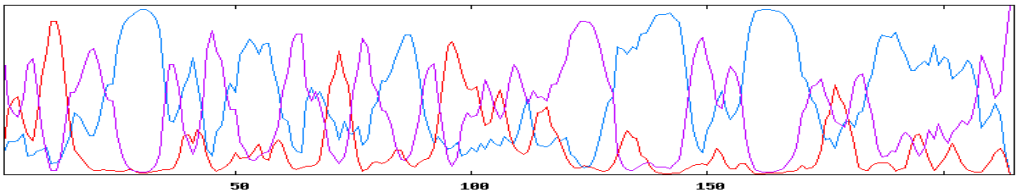
View HNN in: [AnTheProt (PC) , Download...] [HELP]

```
         10         20         30         40         50         60         70
          |          |          |          |          |          |          |
GSHMAARRGALIVLEGVDRAGKSTQSRKLVEALCAAGHRAELLRFPERSTEIGKLLSSYLQKKSDVEDHS
ccceeccccceeeeeeccccccccchhhhhhhhhhhhhcccchhhhccccccchhhhhhhhhhhccccccchhh
VHLLFSANRWEQVPLIKEKLSQGVTLVVDRYAFSGVAFTGAKENFSLDWCKQPDVGLPKPDLVLFLQLQL
eeeeeccccccchhhhhhhhhcccceeeeeeeeeccceeeccccchccccccccccccccccccchhhhhhhhh
ADAAKRGAFGHERYENGAFQERALRCFHQLMKDTTLNWKMVDASKSIEAVHEDIRVLSEDAIATATEKPL
hhhhhhccccccchhcccchhhhhhhhhhhhhhhccccceeecccchhhhhhhhhhhhhhhhhhhhhhhcc
GELWK
ccccc
```

```
Sequence length :   215
```

```
HNN :
   Alpha helix      (Hh) :    96 is  44.65%
   3₁₀ helix        (Gg) :     0 is   0.00%
   Pi helix         (Ii) :     0 is   0.00%
   Beta bridge      (Bb) :     0 is   0.00%
   Extended strand  (Ee) :    27 is  12.56%
   Beta turn        (Tt) :     0 is   0.00%
   Bend region      (Ss) :     0 is   0.00%
   Random coil      (Cc) :    92 is  42.79%
   Ambiguous states (?)  :     0 is   0.00%
   Other states          :     0 is   0.00%
```





SOPMA result for : 1E2Fxx0

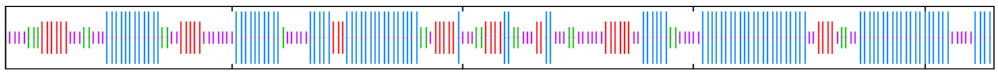
Abstract Geourjon, C. & Deléage, G., SOPMA: Significant improvement in protein secondary structure prediction by consensus predictio

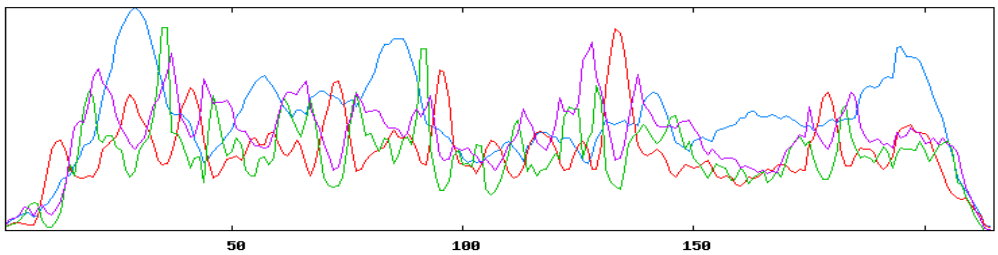
View SOPMA in: [AnTheProt (PC) , Download...] [HELP]

```
         10         20         30         40         50         60         70
          |          |          |          |          |          |          |
GSHMAARRGALIVLEGVDRAGKSTQSRKLVEALCAAGHRAELLRFPERSTEIGKLLSSYLQKKSDVEDHS
hccccctttteeeeeeccctttcccchhhhhhhhhhhhttcceeeeeccccccchhhhhhhhhhtccccchhh
VHLLFSANRWEQVPLIKEKLSQGVTLVVDRYAFSGVAFTGAKENFSLDWCKQPDVGLPKPDLVLFLQLQL
hheeehhhhhhhhhhhhhhhhttceeeeehccctteeeehhttccceehhcccttccccceeeeeeccch
ADAAKRGAFGHERYENGAFQERALRCFHQLMKDTTLNWKMVDASKSIEAVHEDIRVLSEDAIATATEKPL
hhhhhttccccchhhhhhhhhhhhhhhhhhhhhhhhcceeeeecttcchhhhhhhhhhhhhhhhhhhhcccc
GELWK
chhhh
```

```
Sequence length :   215
```

```
SOPMA :
   Alpha helix      (Hh) :   102 is  47.44%
   3₁₀ helix        (Gg) :     0 is   0.00%
   Pi helix         (Ii) :     0 is   0.00%
   Beta bridge      (Bb) :     0 is   0.00%
   Extended strand  (Ee) :    35 is  16.28%
   Beta turn        (Tt) :    20 is   9.30%
   Bend region      (Ss) :     0 is   0.00%
   Random coil      (Cc) :    58 is  26.98%
   Ambiguous states (?)  :     0 is   0.00%
   Other states          :     0 is   0.00%
```

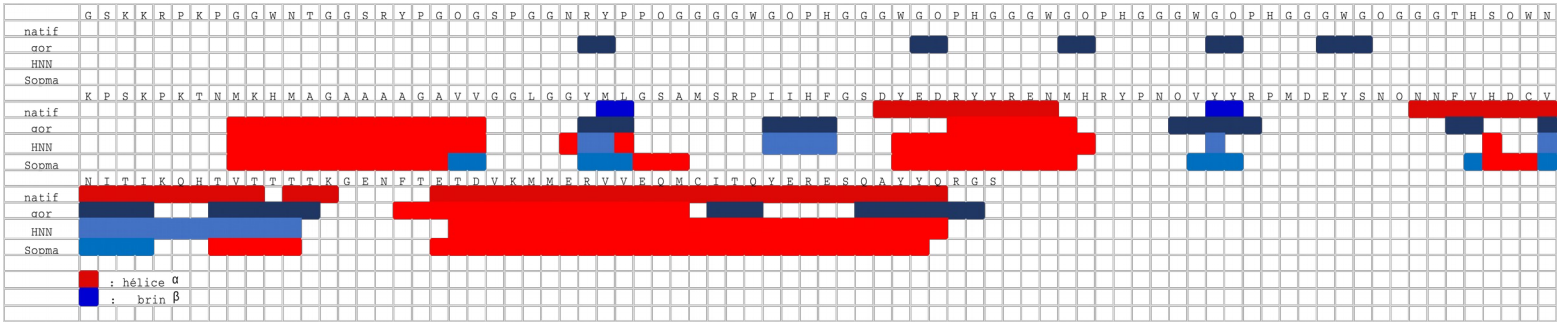




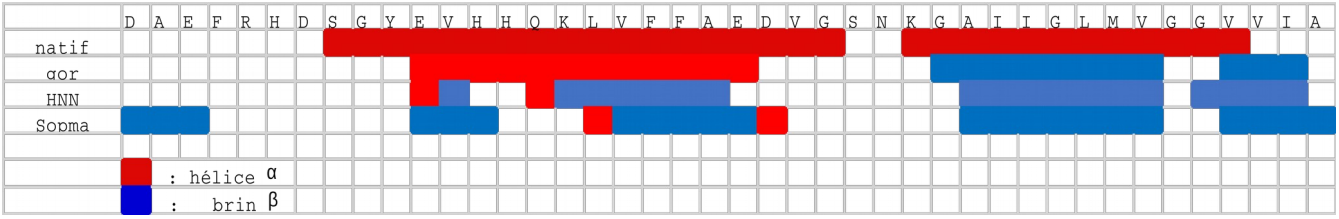
Overall, we notice that the presence of a secondary structure is correctly detected. Nevertheless, the nature of this structure is not always correctly predicted. It's because the only information provided to these programs is the protein sequence. The local interactions (caused by the contiguous amino acids) in the formation of the tertiary structure is neglected and most likely explains the errors. Therefore, rely on only one method seems particularly risky to predict a structure, but alls methods still have the advantage to give an idea of the different possible secondary structures in the protein if we decide to combine.

3. Analysis of two sequences with particular properties

A] Predict the secondary structure of the sequences inconnu1.fasta and inconnu2.fasta (available on the Virtual University), by using the 3 programs used in the previous sections. Record the predictions in the files inconnu1.xls et inconnu2.xls that already contain the secondary structure of the experimental structure of these proteins. Use BLAST (http://www.ncbi.nlm.nih.gov/blast/ ; "Protein Blast") to identify the proteins corresponding to these sequences. Analyse the results, make assumptions and interpret the results.



Then, we used BLAST to identify the structure and this first sequence fits the prion, a disease agent which consists of an unusually folded protein. (https://en.wikipedia.org/wiki/PRNP)



The second one fits a β-amyloïd peptide involved in Alzheimer desease. (https://en.wikipedia.org/wiki/Amyloid_beta)

4. Propensity of amino acids to adopt a given local structure

A] Analyse the propensity of the 20 amino acids to adopt a local structure. Use for that purpose the Fugue program (http://babylone.ulb.ac.be/Prelude_and_Fugue/Fugue/index.php). Submit to this program 20 peptides composed by 20 amino acids of the same type (AAAAAAAAAAAAAAAAAAAA, ...). Read carefully the description of Fugue and how to interpret its results (http://babylone.ulb.ac.be/Prelude_and_Fugue/). You can divide this work in groups.

Intro descriptive de Prelude&Fugue: PRELUDE & FUGUE

predict the local structure of a protein in terms of backbone dihedral angle domains, identify sequence regions that form early during folding, and locate structural weaknesses, defined as

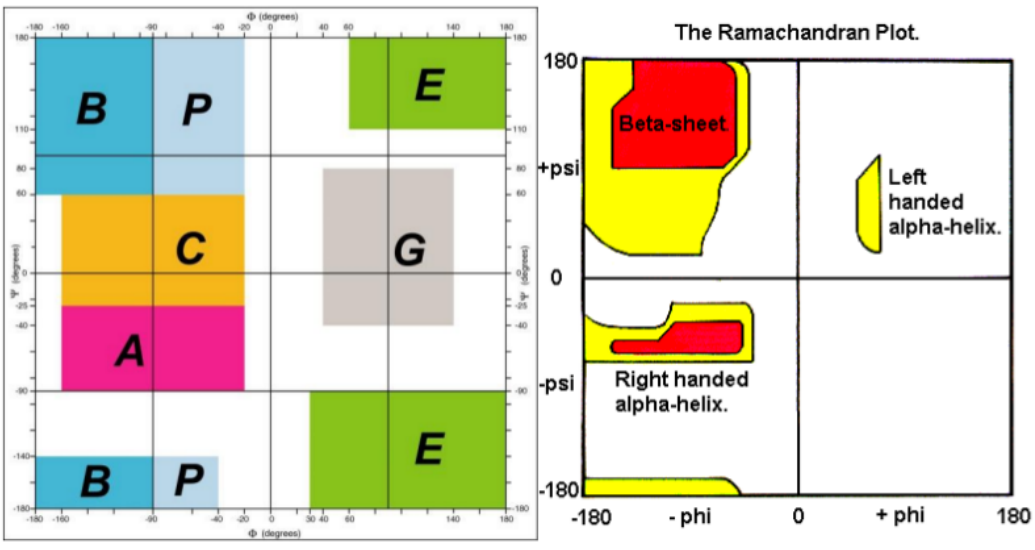
regions whose sequence is not optimal with respect to the tertiary fold.
These programs use a statistical backbone torsion angle potential, which describes local interactions along the chain and is derived from a set of 1403 known protein structures. The input they require is the amino acid sequence of the target protein. The output yields the predicted structures given in terms of 7 backbone torsion angle assignments noted A, B, C, G, E, O, P; the phi, psi, and omega values associated to these assignments are given here.

Prelude & Fugue predicts the N backbone conformations of lowest energy of a protein sequence or sequence segment, where N is specified by the user. The user may impose constraints on some interatomic distances in the predicted structures. The output file contains the lowest energy structures that satisfy the constraints, the predicted energy values, and the energy gap and root mean square (rms) deviation of superimposed backbone atoms of each predicted structure relative to the lowest energy predicted structure. In addition, the 3D structures of all the predicted conformations are supplied in PDB format.

A sequence whose lowest energy conformation displays a sizable energy gap relative to other predicted structures is considered to have a well defined preferred conformation. The program is designed to run mainly on short peptides, because it only considers local interactions along the chain and overlooks tertiary interactions. When applied to longer peptides, the predictions must be considered as 2D rather than 3D and similar to secondary structure predictions.

Prelude & Fugue predicts the backbone structure, not of the whole input sequence, but of those segments whose lowest energy structure is strongly preferred over other conformations. The strength of the prediction is given by a weight between 1 and 9.

Regions predicted by Fugue, with high weights, have an strong intrinsic preference for their predicted conformation in the absence of tertiary interactions. When cut from the sequence, the peptides so obtained are likely to adopt preferentially this conformation in water at low temperature or in an apolar medium. When considered within the sequence, these regions can be exepected to form at the very beginning of the folding process. Regions where the predicted structure differs from the native structure are likely to be regions whose instrinsic structural preference, determined by local interactions along the chain, are modified due to tertiary interactions. They are interpreted as structural weaknesses, which possibly slow down folding and cause alternative structuring.



The comparison between the Ramachandran diagram for localization according to the angles Φ and Ψ and the Prelude & Fugue technique allows to highlight two groups of interesting amino acids: The amino acids present in zone B will therefore tend to be present in the region of the β sheets and the amino acids present in region A will tend to end up in the α helices:
- α : Alanine, Histidine, Lysine, Arginine, Glutamine ;
- B : Valine, Isoleucine

B] Use the results obtained in section A to test several mutations (single-site and multiple) in the sequence inconnu2.fasta, with the aim of increasing the propensity of the second helix to be predicted as an helix. Use the HNN program to make the predictions on the mutated sequences. Analyse the results. Is it possible to increase the propensity to adopt an helix by mutating some sequence positions?

If we modify the amino acids Valine and Isoleucine with the amino acids present mainly in the helices, we can obtain a more oriented prediction. it is not necessary to introduce a large number of mutations to modify the intrinsic equilibrium of the protein.