



Metodi predittivi 1D

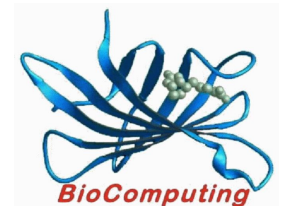
BioComputing


Introduzione

Data la sequenza di una proteina ci sono tutta una serie di predizioni che si possono fare. Queste utilizzano i metodi di predizione che abbiamo visto in precedenza, in particolare i metodi di *machine learning*.

Un elenco di predittori che vedremo:

- 1) *Struttura secondaria*
- 2) *Accessibilità*
- 3) *Transmembrana*
- 4) *Coiled-coil*
- 5) *Disordine*
- 6) *Repeats*
- 7) *Peptidi segnale*

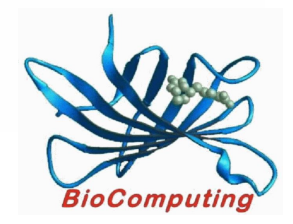
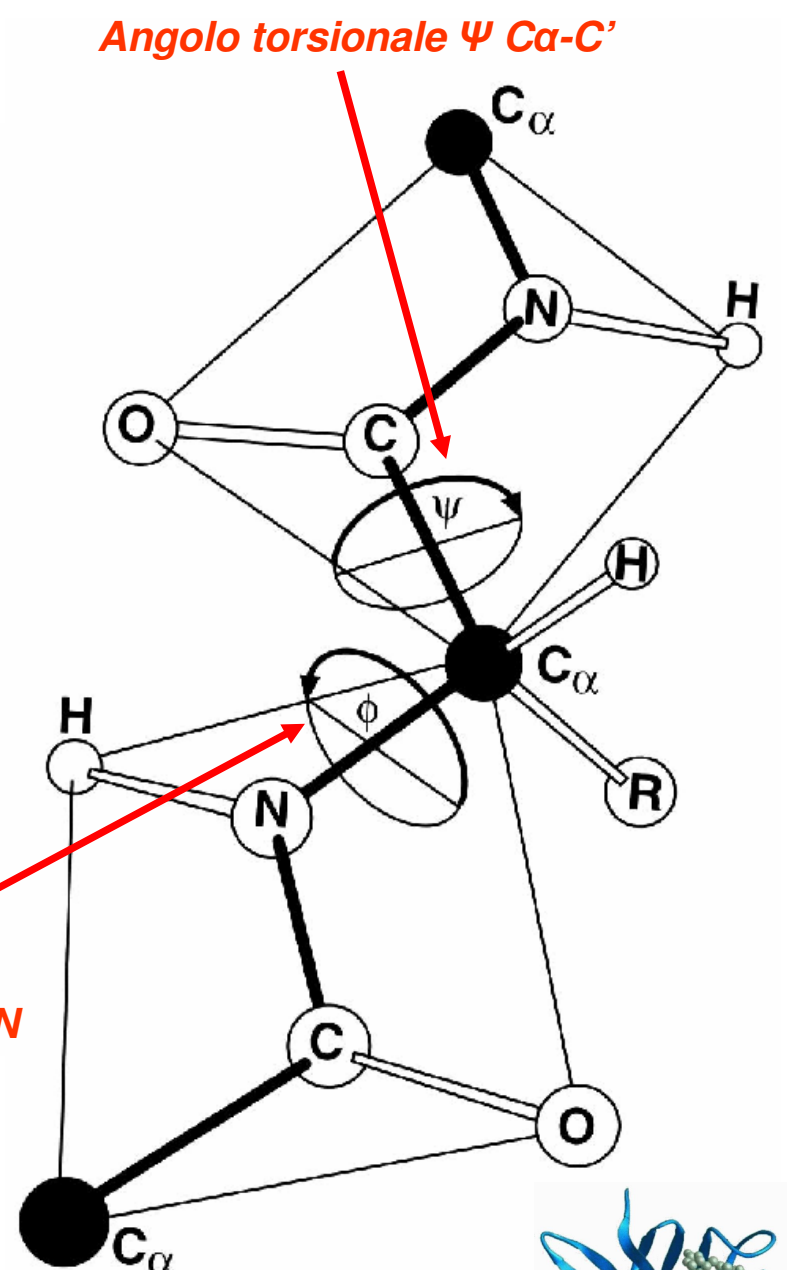
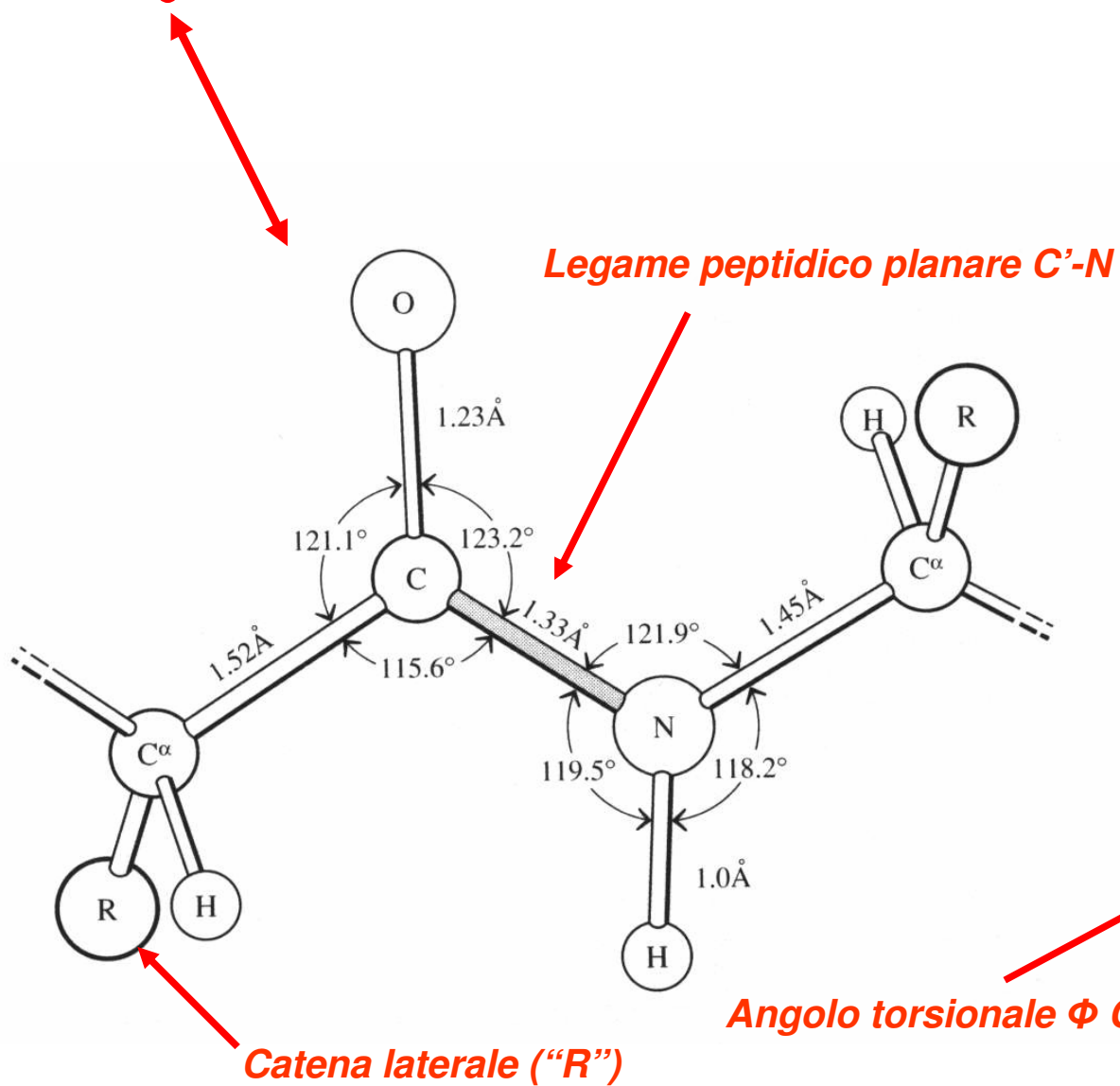


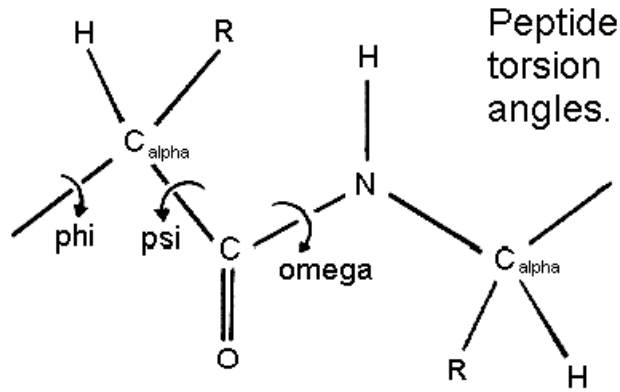


***STRUTTURA
SECONDARIA***

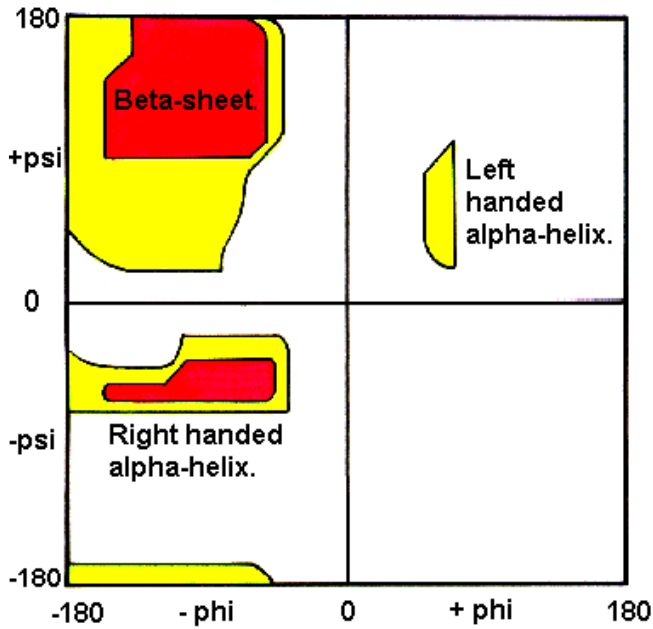
BioComputing

NKSKFCANAILAVSLANAK.AAAA.AKGMPLYEHIAELNGTPGKYSMPVPM





The Ramachandran Plot.



STRUTTURA SECONDARIA

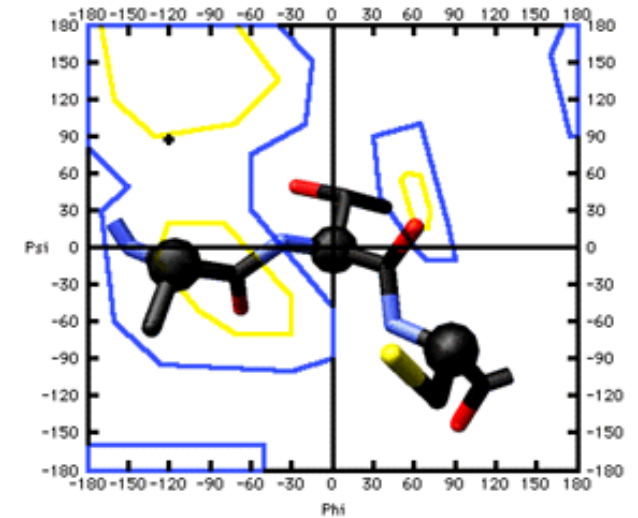
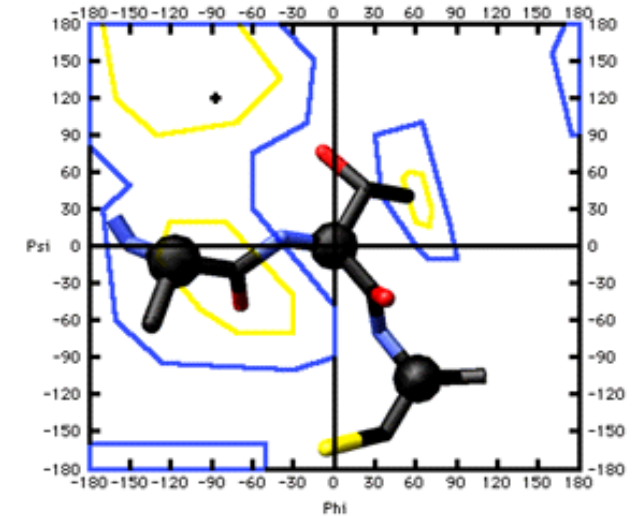
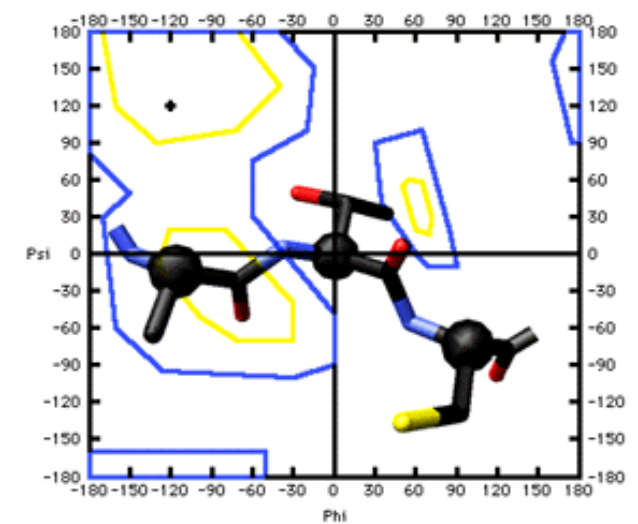
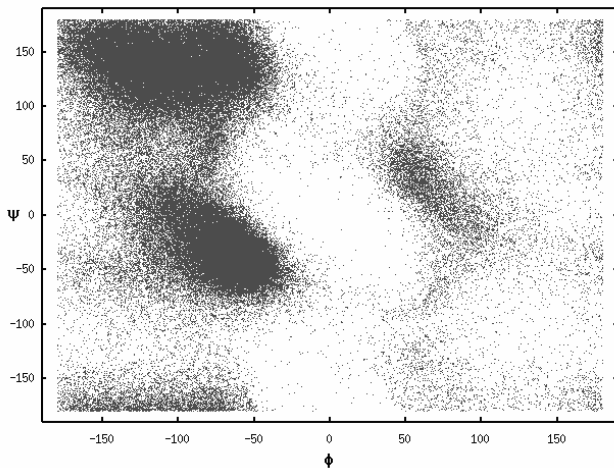
Legami N-C α phi (Φ) e C α -C psi (Ψ) liberi di ruotare

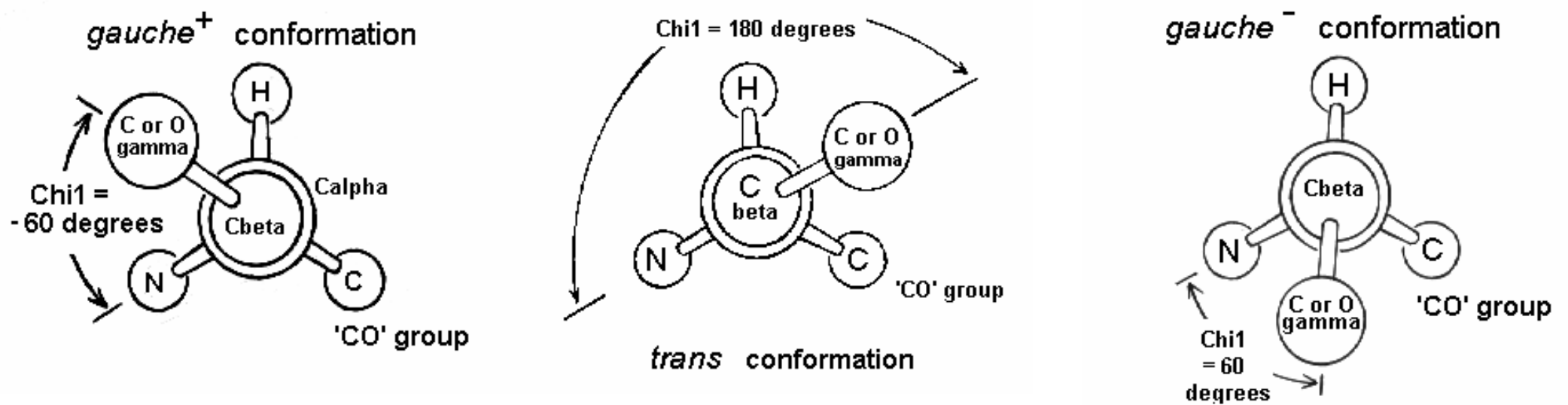
Causa planarit  legame peptidico Omega (Ω) 180 $^\circ$ trans. Cis pi  rara a 0 $^\circ$ coinvolge la prolina

Ramachandran plot

legami phi e psi dei legami peptidici. Analisi delle collisioni degli atomi considerando il raggio di van der Waals. Zona rossa nessuna collisione. Zona gialla al limite della collisione. Zona bianca collisione tra atomi se avessero legami peptidici con certi phi e psi.

Elica 3_{10}   instabile

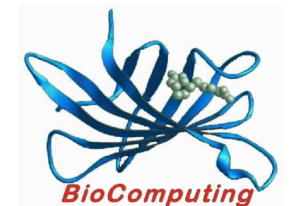




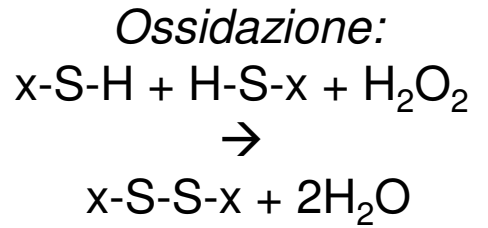
Le 20 catene laterali

Amino acid or residue thereof	Three-letter symbol	One letter symbol	Mnemonic help for one-letter symbol	Relative abundance in <i>E. coli</i> proteins (19) (%)	M.W. of residue at pH7.0 (daltons)	pK value of side chain (19)	ΔG values for transfer of side chain from water to ethanol at 25°C (16) (kcal/mol)
Alanine	Ala	A	<u>A</u> lanine	13.0	71		-0.5
Glutamate	Glu	E	glu <u>E</u> tamic acid		128	4.3	
Glutamine	Gln	Q	<u>Q</u> -tamine	10.8	128		
Aspartate	Asp	D	aspar <u>D</u> ic acid	9.9	114	3.9	
Asparagine	Asn	N	asparagi <u>N</u> e		114		
Leucine	Leu	L	<u>L</u> eucine	7.8	113		-1.8
Glycine	Gly	G	<u>G</u> lycine	7.8	57		
Lysine	Lys	K	before <u>L</u>	7.0	129	10.5	
Serine	Ser	S	<u>S</u> erine	6.0	87		+0.3
Valine	Val	V	<u>V</u> aline	6.0	99		-1.5
Arginine	Arg	R	a <u>R</u> ginine	5.3	157	12.5	
Threonine	Thr	T	<u>T</u> hreonine	4.6	101		-0.4
Proline	Pro	P	<u>P</u> roline	4.6	97		
Isoleucine	Ile	I	<u>I</u> soleucine	4.4	113		
Methionine	Met	M	<u>M</u> ethionine	3.8	131		-1.3
Phenylalanine	Phe	F	<u>F</u> enylalanine	3.3	147		-2.5
Tyrosine	Tyr	Y	t <u>Y</u> rosine	2.2	163	10.1	-2.3
Cysteine	Cys	C	<u>C</u> ysteine	1.8	103		
Tryptophan	Trp	W	t <u>W</u> o rings	1.0	186		-3.4
Histidine	His	H	<u>H</u> istidine	0.7	137	6.0	-0.5

Weighted mean 108.7

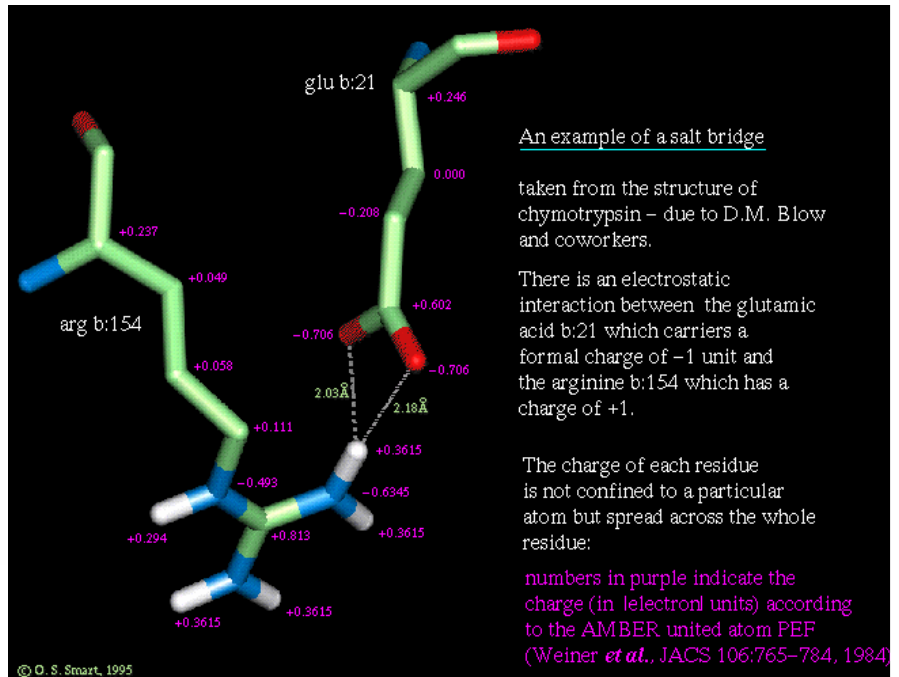
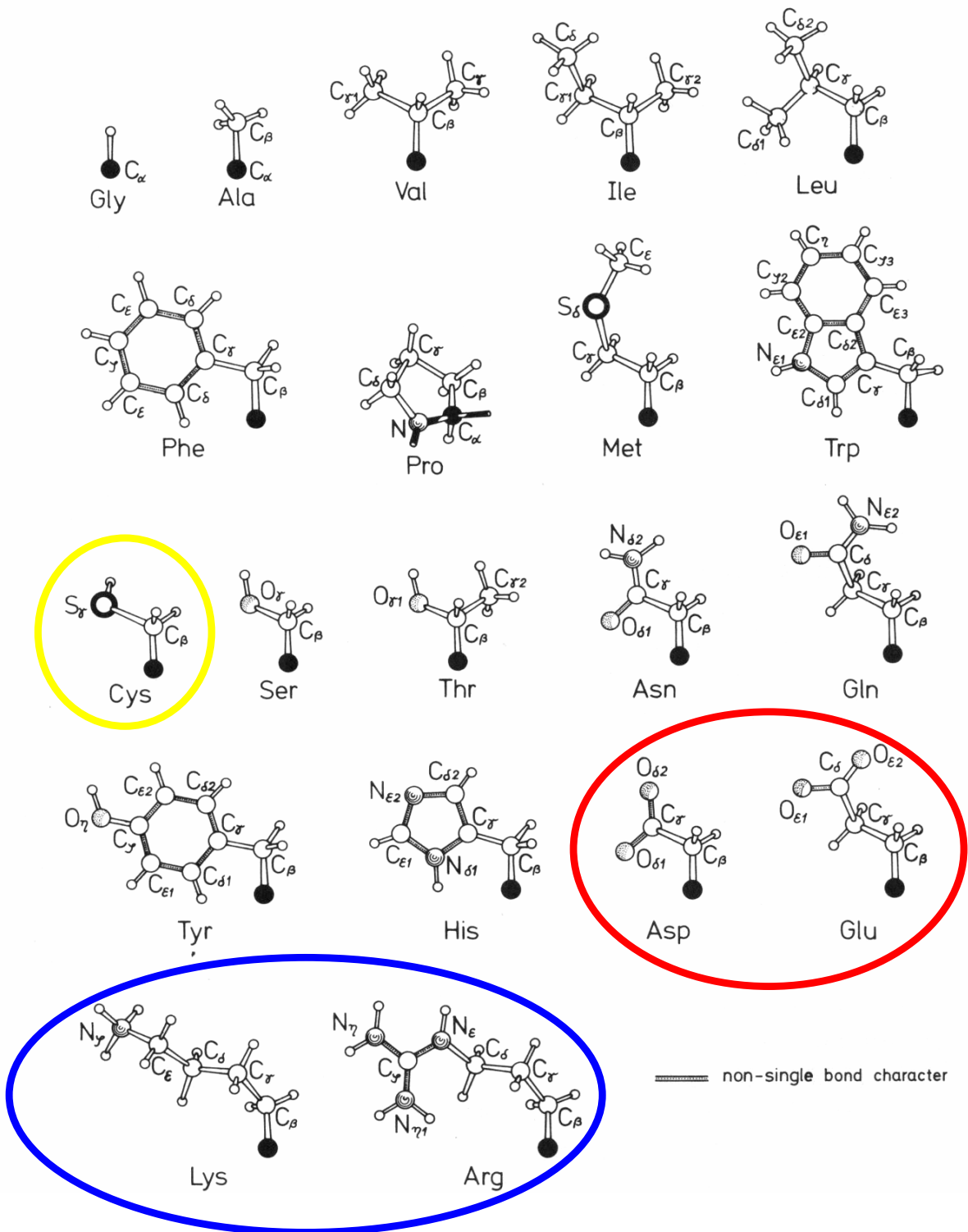


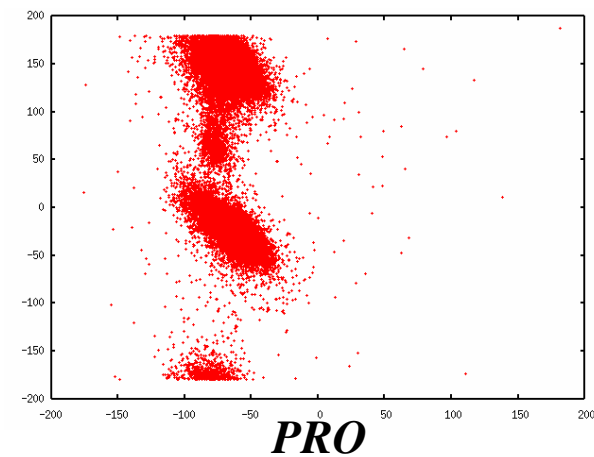
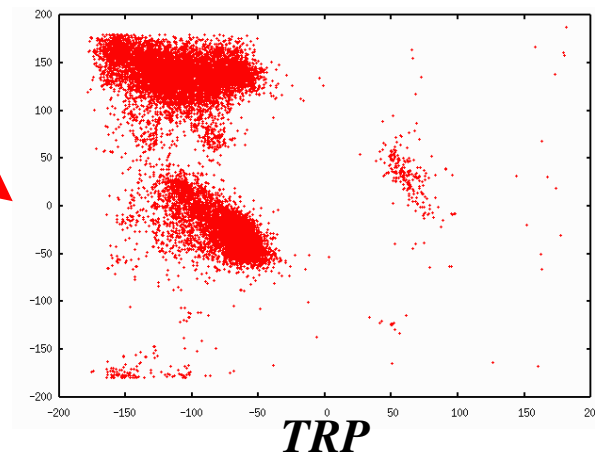
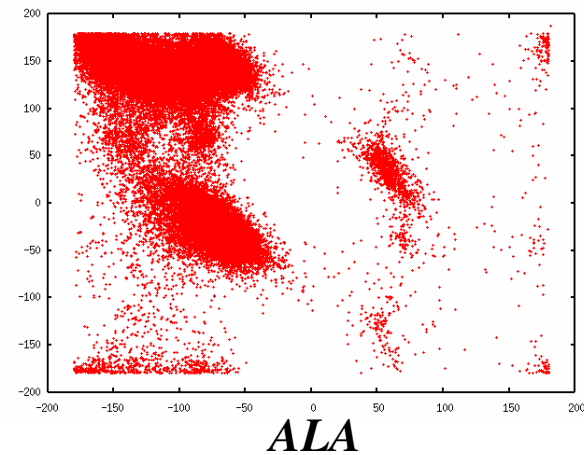
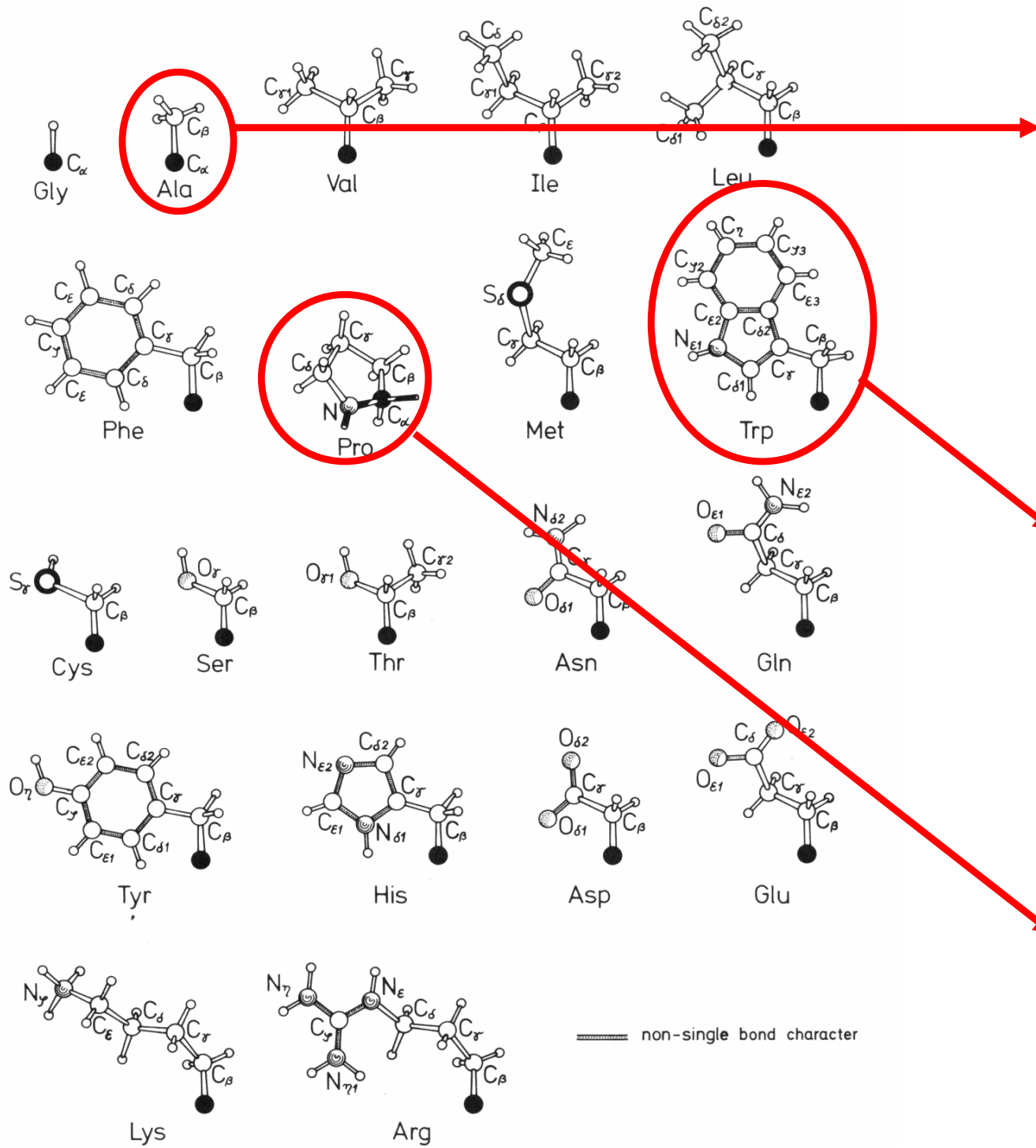
Può formare disolfuri

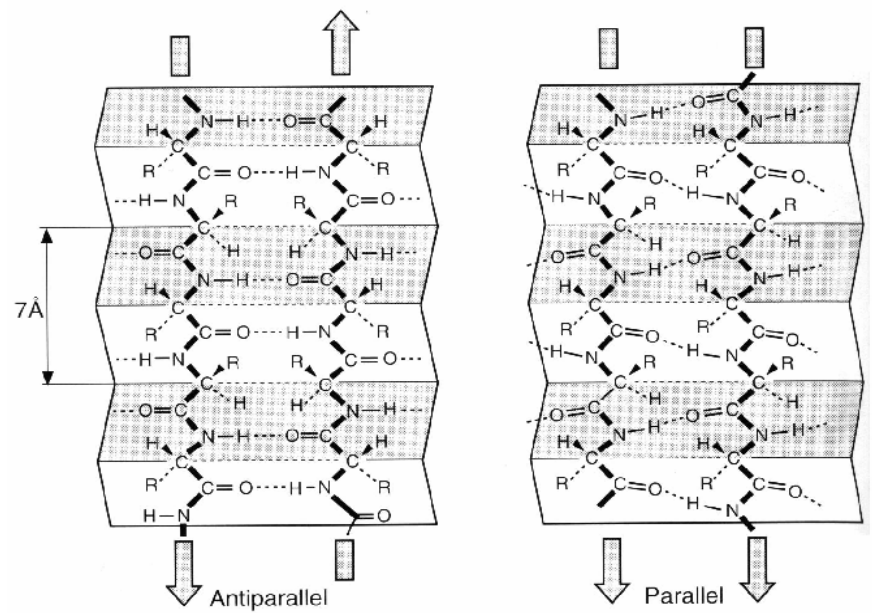
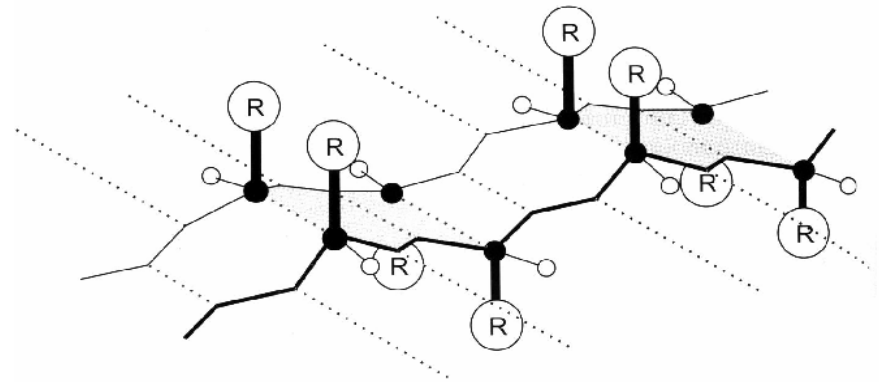
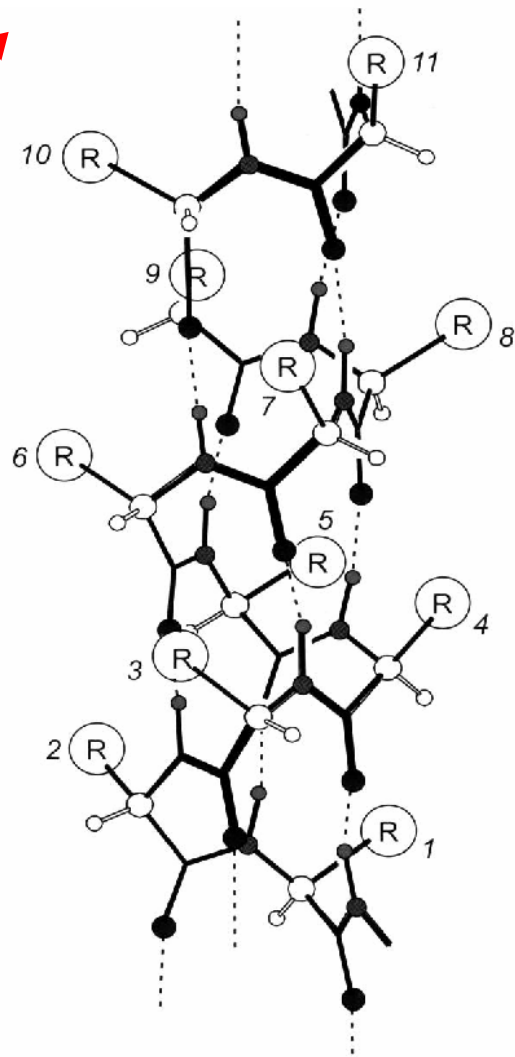
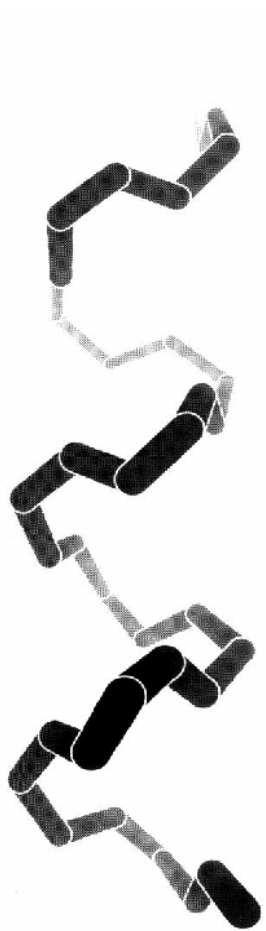
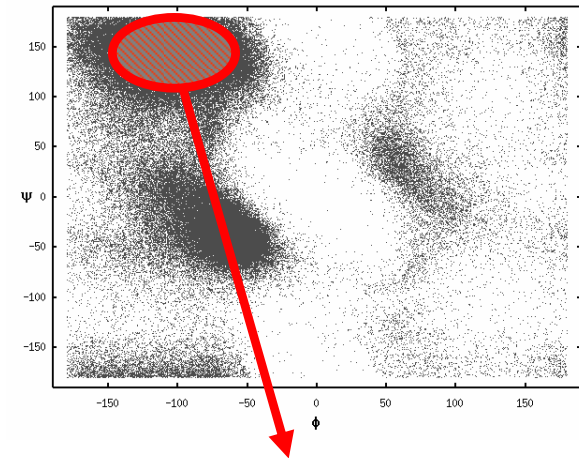
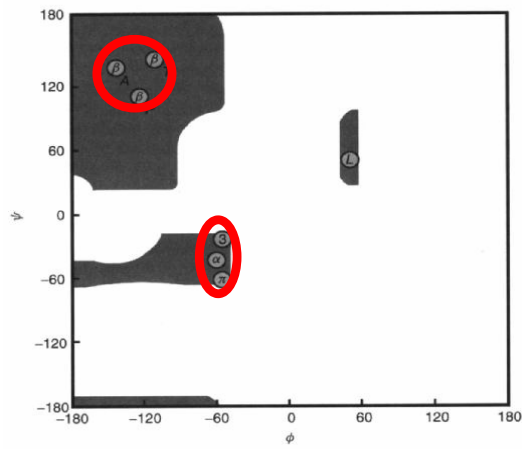
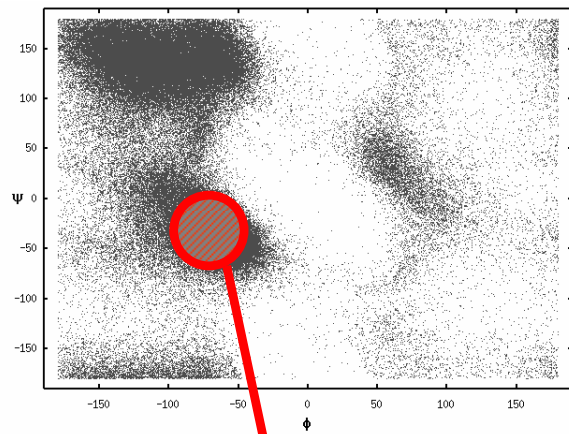


Carica negativa

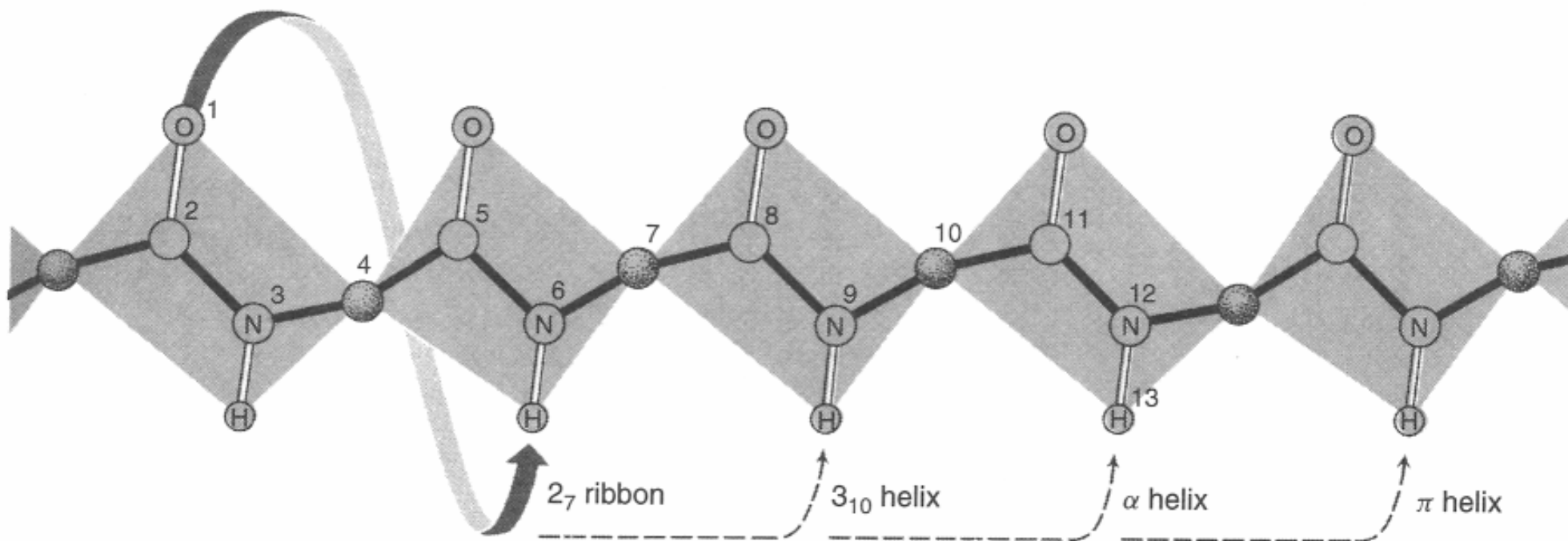
Carica positiva



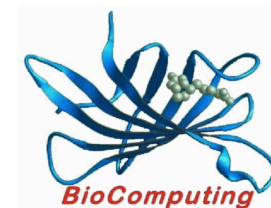




CONVENZIONI PER ALFA ELICA

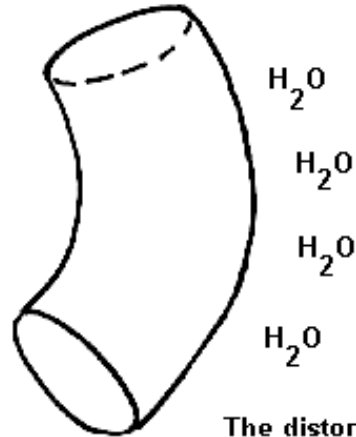


$(i \rightarrow i+4)$	alpha elica sito accettore	$(>)$	$O \rightarrow H-N$
$(i \rightarrow i+3)$	elica 3-10 sito accettore	$(>)$	$O \rightarrow H-N$
$(i-4 \rightarrow i)$	alpha elica sito donatore	$(<)$	$N-H \rightarrow O$
$(i-3 \rightarrow i)$	elica 3-10 sito donatore	$(<)$	$N-H \rightarrow O$

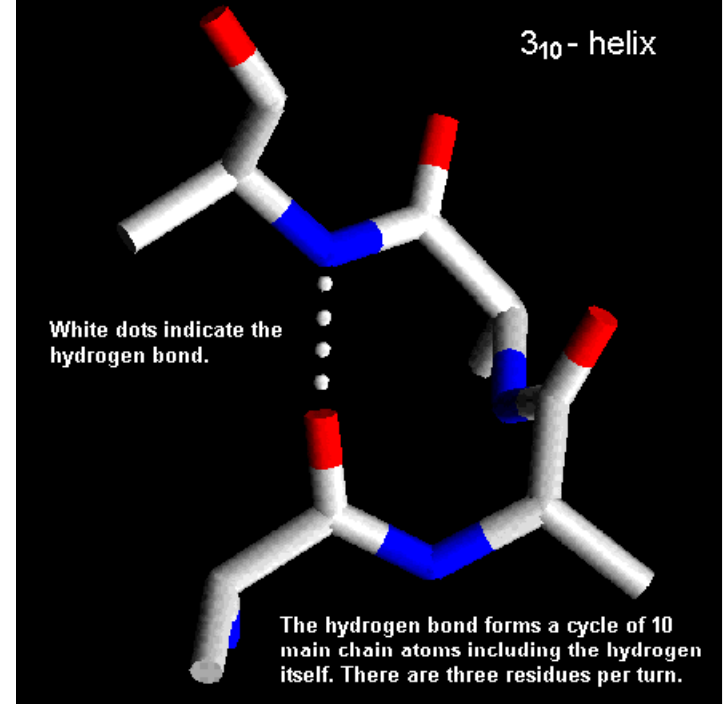


α -elica

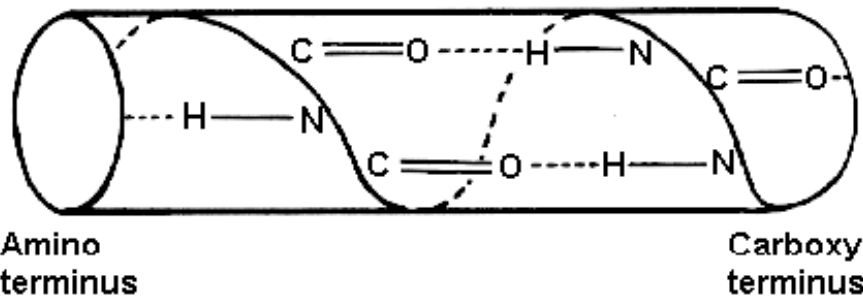
Solvent induced distortion of an alpha helix.



The distortion is exaggerated for clarity.



Toilet roll representation of the main chain hydrogen bonding in an alpha-helix.

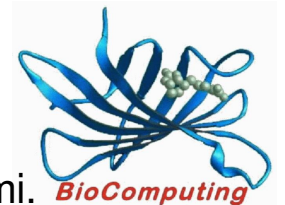


Struttura si ripete ogni 5.4 Å e 3.6 AA. Ogni AA dà una altezza di 1.5 Å all'elica. È right handed C=O e N-H ogni 4 AA

Angoli phi e psi negativi, tipicamente -60 e -50 rispettivamente

Distorsione a causa della prolina rigida che rompe i ponti-H e le parti esposte al solvente. C=O tende a formare pont-H con acqua

Elica 3₁₀ è rara in genere alla fine dell' alpha-elica e ponte-H ogni 3 AA e forma anello a 10 atomi.



β -strand

Angoli Φ sono negativi e Ψ sono positivi.
 Tipicamente $\Phi = -140$ e $\Psi = 130$

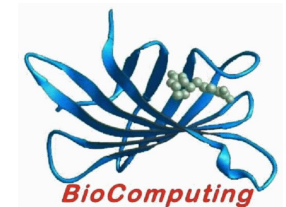
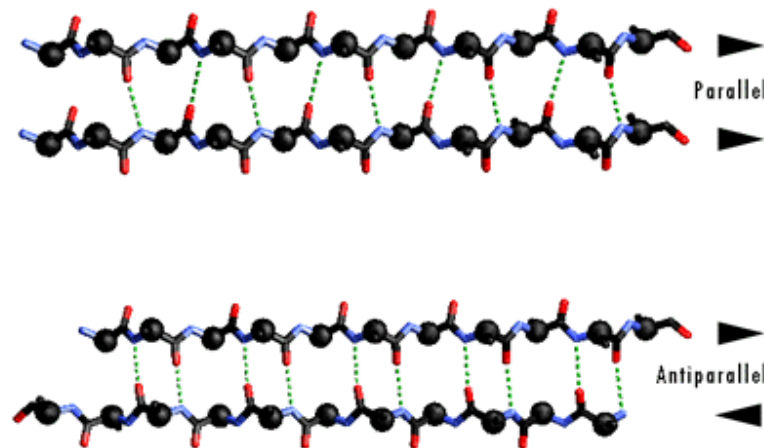
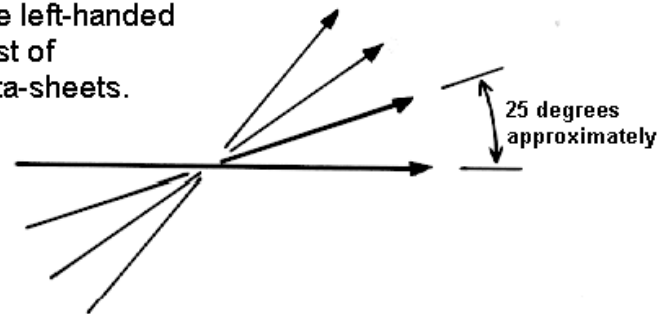
Conformazione a zig-zag

Subunità ripetute a 2 AA di altezza complessiva 7 Å. Ogni AA contribuisce per 3.5 Å

Paralleli più instabili. Ponti-H distorti in tensione. Sempre interni alla catena polipeptidica. Non esposti

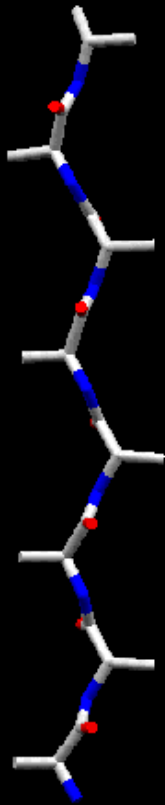
Antiparalleli più stabili e più esposti in superficie

The left-handed twist of beta-sheets.



A diagram of a polypeptide in the beta conformation.

Note the pronounced zig-zag appearance.

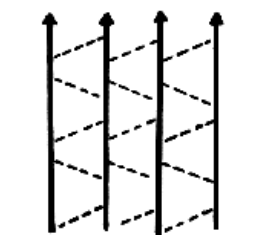
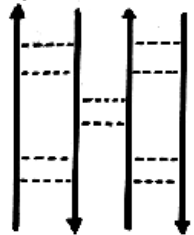


The peptide bonds of adjacent residues point in opposite directions towards and away from the plane of the screen.

Alternate side chains also point in opposite directions approximately in the plane of the screen.

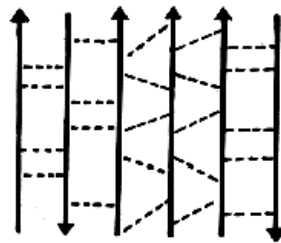
Leggera rotazione sull'asse del piano

Antiparallel beta-sheet



Parallel beta-sheet

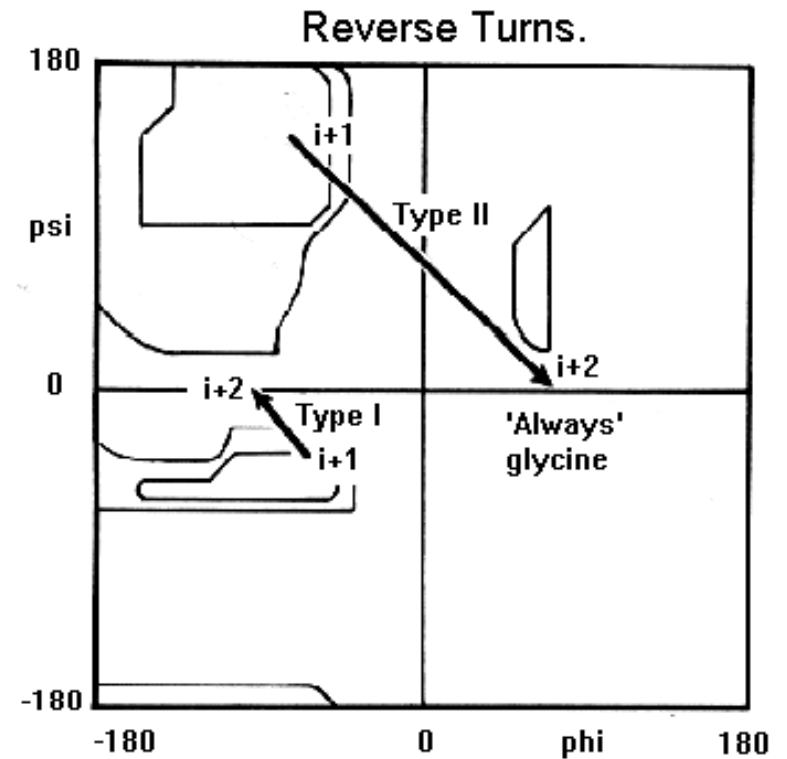
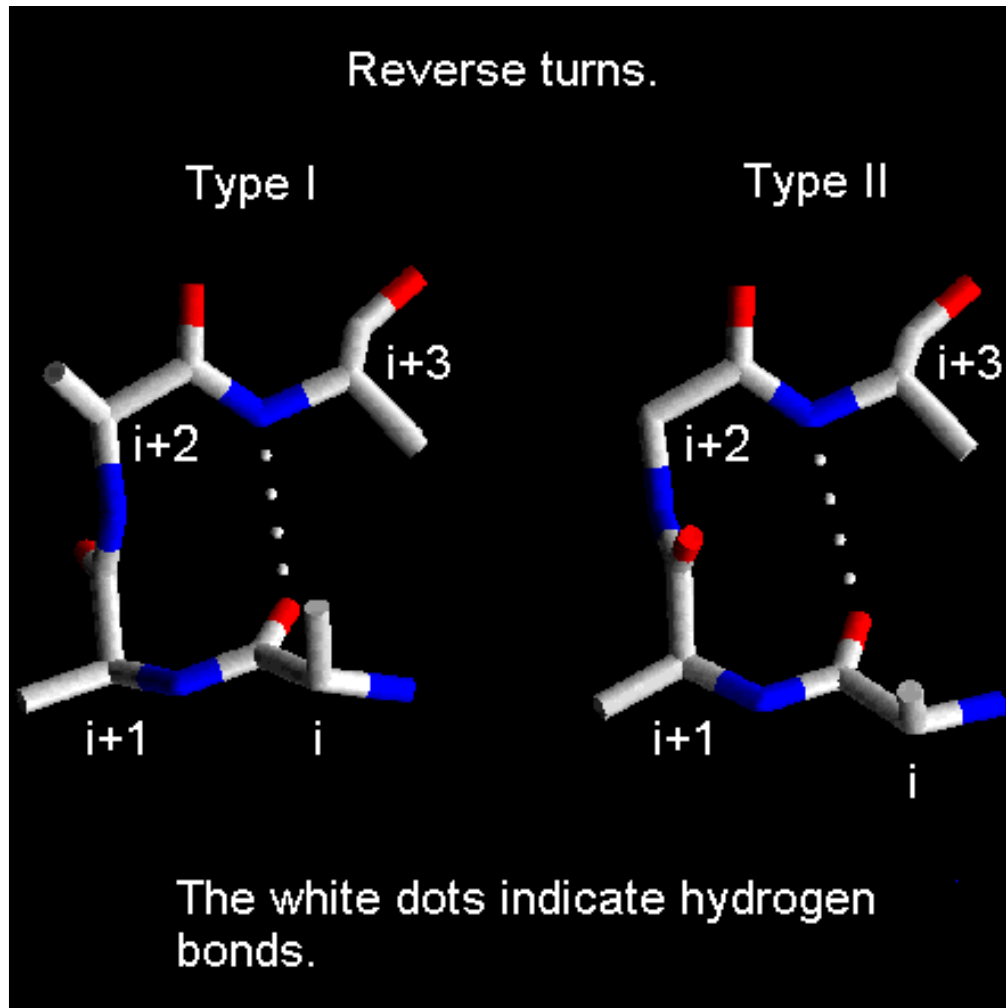
The different types of beta-sheet. Dashed lines indicate main chain hydrogen bonds.



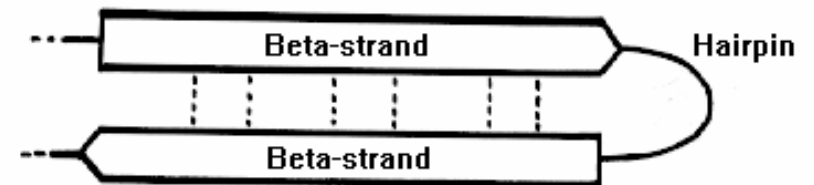
Mixed beta-sheet

I Reverse turns

Conformazione particolare di ponte-H tra C=O e N-H a 3 AA di distanza. Non è α -elica né β -hairpin. Sono abbondanti in proteine globulari ed in superficie. Centri di nucleazione del folding (?). Tipo I e II



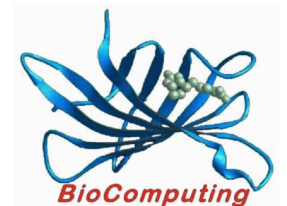
The beta-hairpin turn.



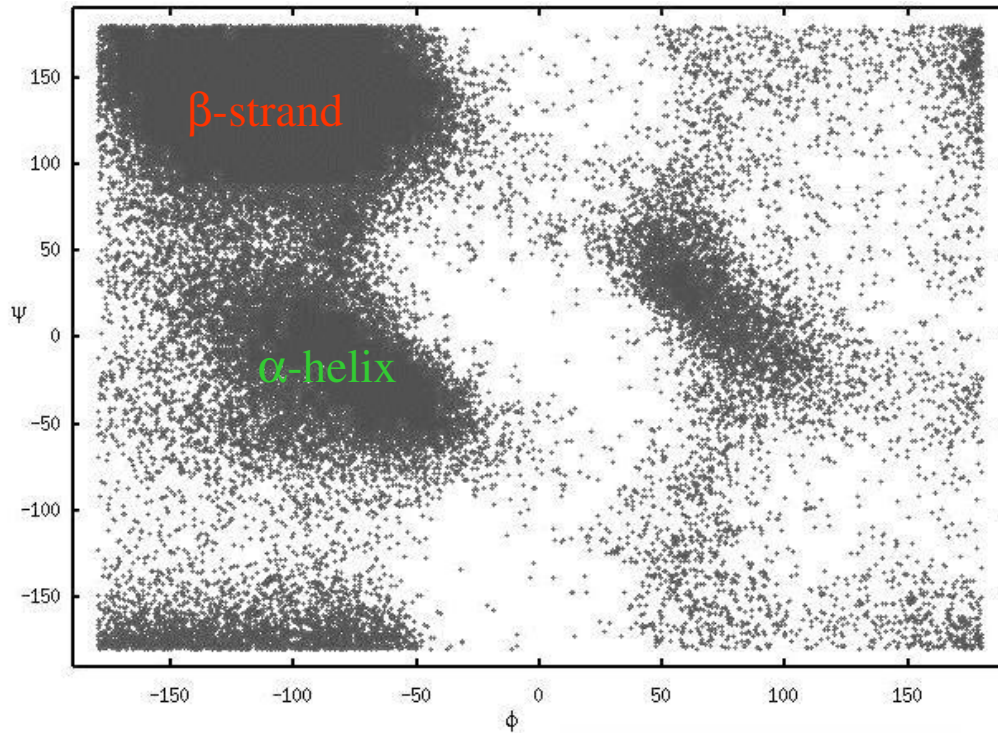
The dashed lines indicate main chain hydrogen bonds.

STRUTTURA SECONDARIA DELLE PROTEINE

- 1) Definizione di struttura secondaria tramite metodi che analizzano i file di PDB attraverso le coordinate spaziali degli AA e i loro angoli torsionali
- 2) Metodi predittivi basati sulla propensione degli AA a certe strutture secondarie (Chou Fasman)
- 3) Metodi predittivi evoluti basati su reti neurali
- 4) I metodi consenso



PREDIZIONE STRUTTURA SECONDARIA DAI FILE PDB



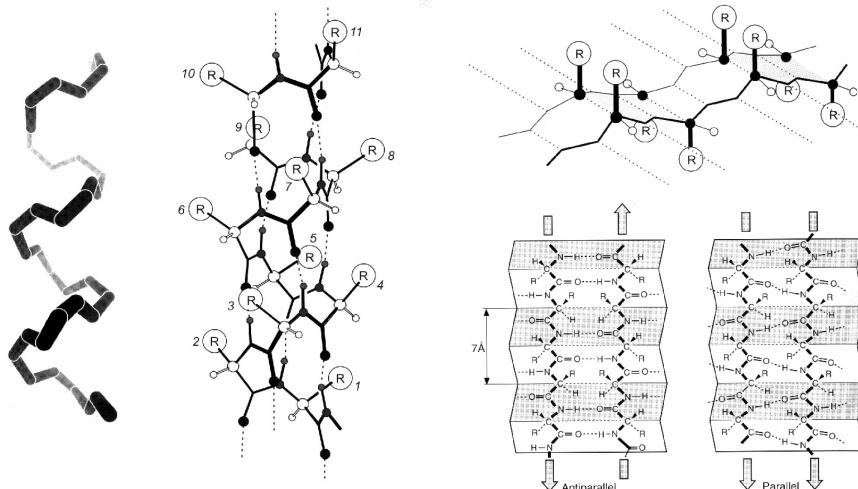
- Predire la struttura secondaria di una proteina è un primo passo comunemente utilizzato per la sua classificazione ed il modelling.

- Tre stati sono generalmente predetti:

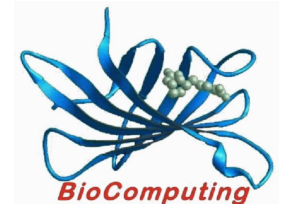
α -elica ('H')

filamento β ('E', per 'extended')

coil/loop ('C' o ':')



Come si determina la struttura secondaria nei file PDB?



Struttura secondaria - Modelli di ponti idrogeno

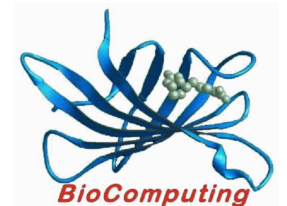
Concetto:

Nelle proteine gli AA idrofobici si trovano all'interno e minimizzano i contatti con il solvente che è polare. Si pensa sia la forza principale per il folding delle proteine.

Le strutture secondarie specifiche sono invece governate da ponti-H intraproteina (*Hvidt and Wesh, 1998*). Ovvero il ripiegamento all'interno degli AA idrofobici porta all'interno anche quelli polari della backbone.

L'effetto porta ad un aumento dell'energia, quindi devono essere bilanciate le forze in gioco e i ponti-H sono tra questi. Se tutte gli AA del core fossero idrofobici le forze di interazioni sarebbero deboli per tenere il core stabile.

- 90% dei gruppi C=O e N-H hanno ponti-H (*Baker-Hubbard, 1984*)
- 62% hanno ponti-H intra backbone (*Andersen 2001*)



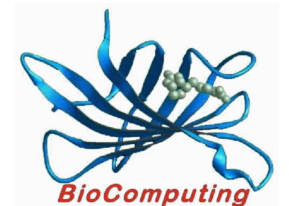
CRITERI PER DETERMINARE UN PONTE-H

Cos'è un ponte di idrogeno ragionevole?

I criteri utilizzati per deciderlo sono un tanti e non sempre concordano.

- ***Geometria***
- ***Potenziale elettrostatico***

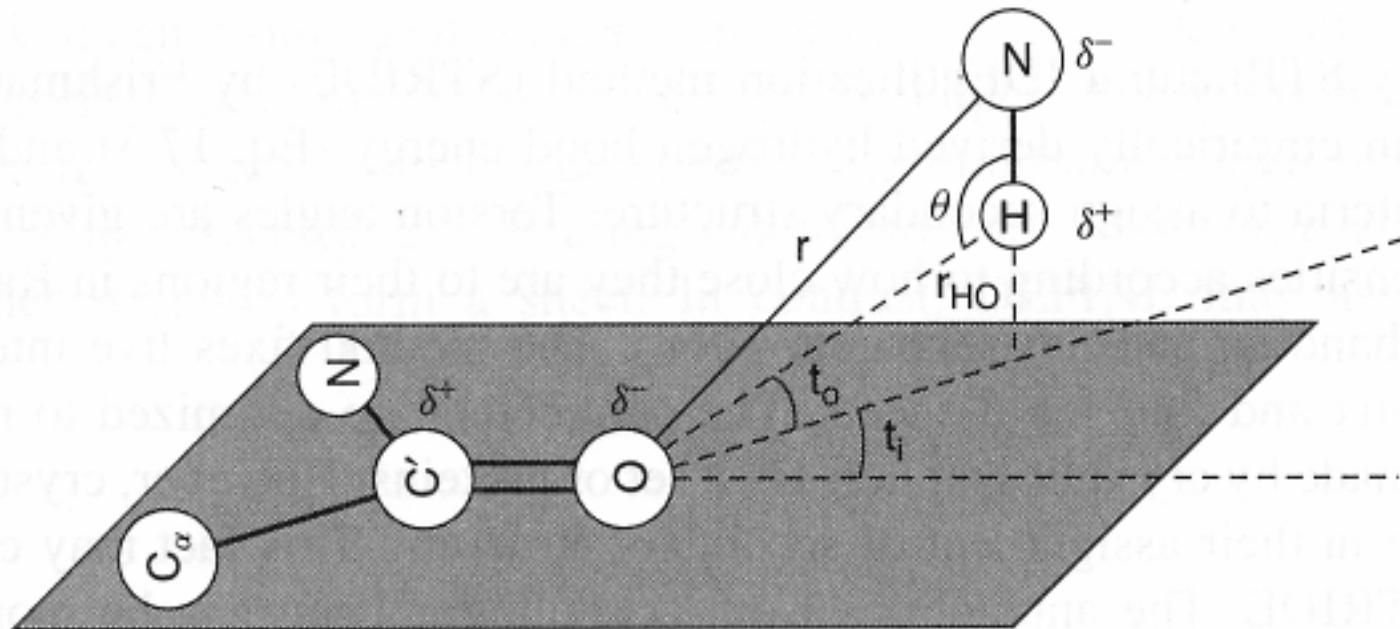
NB: I criteri sopraelencati sono solamente applicabili se le posizioni degli atomi di idrogeno sono note. Non tutte le strutture cristallografiche li possiedono, però è possibile ricostruirli.



CRITERIO GEOMETRICO

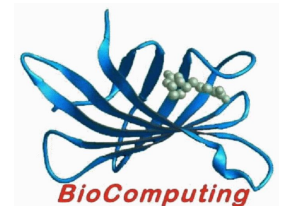
Definizione di ponti-H (*Baker-Hubbard, 1984*):

Angolo compreso tra $N \rightarrow H \rightarrow O = \theta > 120^\circ$ e la distanza r_{HO} 2.5-3.5 Å



Identificazione di struttura secondaria basata sul calcolo dell'energia

- Come possiamo definire dove inizia e finisce un elemento di struttura secondaria?
- Un problema “triviale ma difficile” (*Richardson, 1981*).
- Non esiste un singolo algoritmo corretto per assegnare la struttura secondaria.
- Generalmente si usano criteri come la **conformazione della backbone** (ϕ, ψ) e pattern di **legami di idrogeno**.
- **DSSP** (*Kabsch & Sander, 1983*) e **STRIDE** (*Frishman & Argos, 1995*) sono i due programmi comunemente usati, anche se esistono molti modi di definire i bordi della struttura secondaria.



CRITERI BASATI SUL CALCOLO DELL'ENERGIA

1) **DSSP** calcolo energia elettrostatica e pattern di ponti-H.

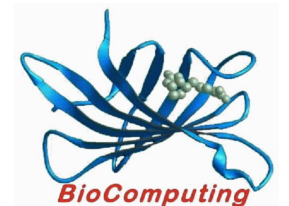
2) **STRIDE** calcolo energia diverso da DSSP e pattern ponti-H simile a DSSP. Valutazione degli angoli torsionali.

I programmi STRIDE e DSSP come anche altri (P-CURVE, ecc.) definiscono varie topologie di struttura secondaria che si possono far ricondurre a tre classi fondamentali:

C (*coil*) sono le regioni non ordinate o predette oppure i turn ed i beta-hairpin.

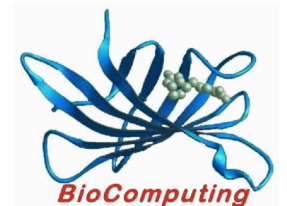
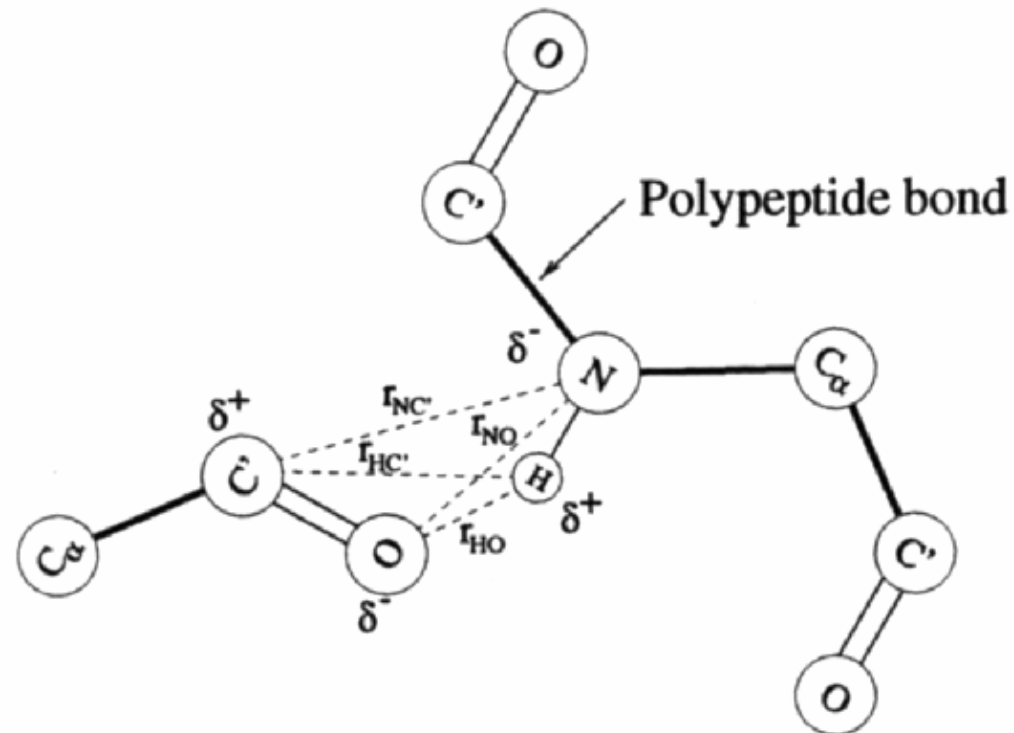
H (*helix*) sono tutti i tipi di configurazione ad elica.

E (*extended*) sono i foglietti beta.



DSSP

- Dictionary of secondary structure in proteins (*Kabsch & Sander, 1983*)
- *DSSP* cerca ponti di idrogeno per assegnare α -eliche e filamenti β .
 - La definizione può essere ambigua.



Calcolo dell'Energia del legame ponte-H basato sulla legge di Coulomb
 Una alternativa all'assegnamento del ponte-H è il calcolo dell'energia di Coulomb basata sull'attrazione elettrostatica utilizzata da DSSP.

$$E = f\delta^+\delta^- \left(\frac{1}{r_{NO}} + \frac{1}{r_{HC'}} + \frac{1}{r_{HO}} + \frac{1}{r_{NC'}} \right)$$

f è una costante

δ^+ e δ^- cariche polari date in unità della carica elementare dell'elettrone

Cutoff per definire che un legame è un ponte-H si basa su una $E < -0.5$ kcal/mol

Non calcola la repulsione tra atomi e non impone delle lunghezze dei ponti-H caratteristici.

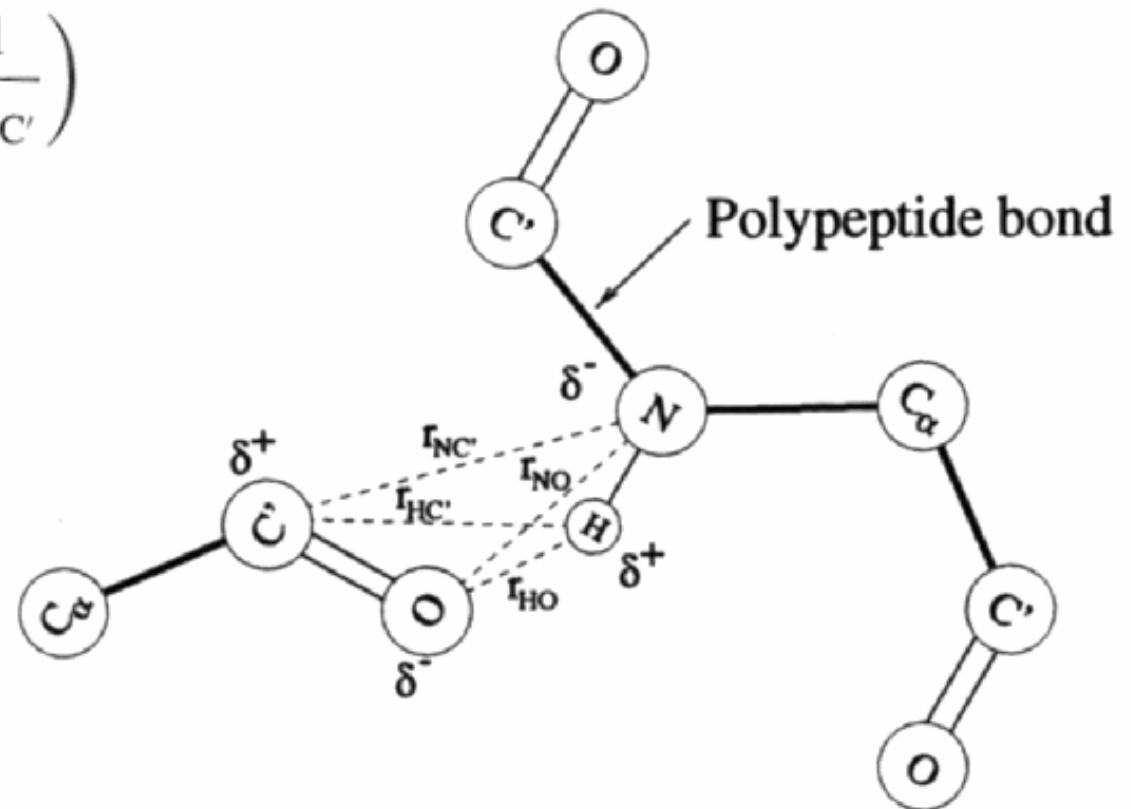
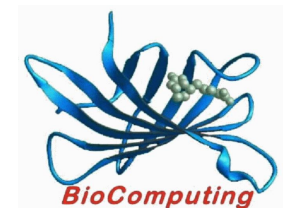


Fig. 1. The distances used to calculate the Coulomb H bond.



STRIDE

STRuctural IDentification (*Frishman e Argos 1995*). Calcolo dell'energia del ponte-H empirico ricavato dalla geometria e dall'energia di distanza N-O (calcola l'eventuale repulsione se N-O sono troppo vicini) e dai tre angoli di legame.

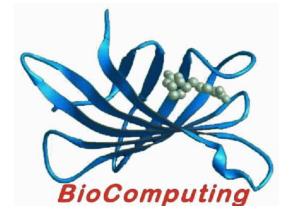
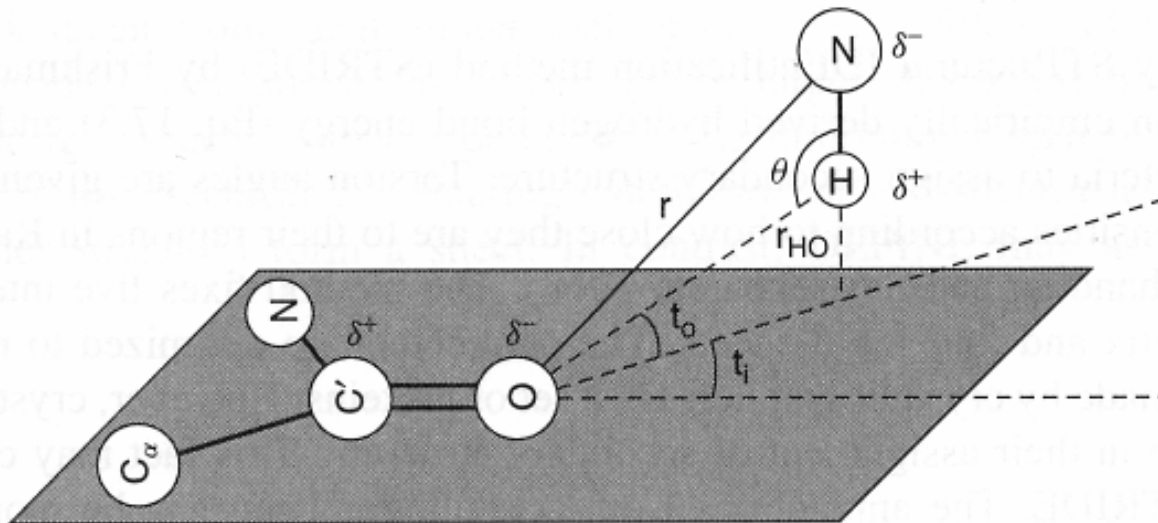
E_r energia ricavata dalla distanza di N-O.

E_t e E_p energie empiriche basate sui tre angoli della figura.

$$E_{hb} = E_r \cdot E_t \cdot E_p$$

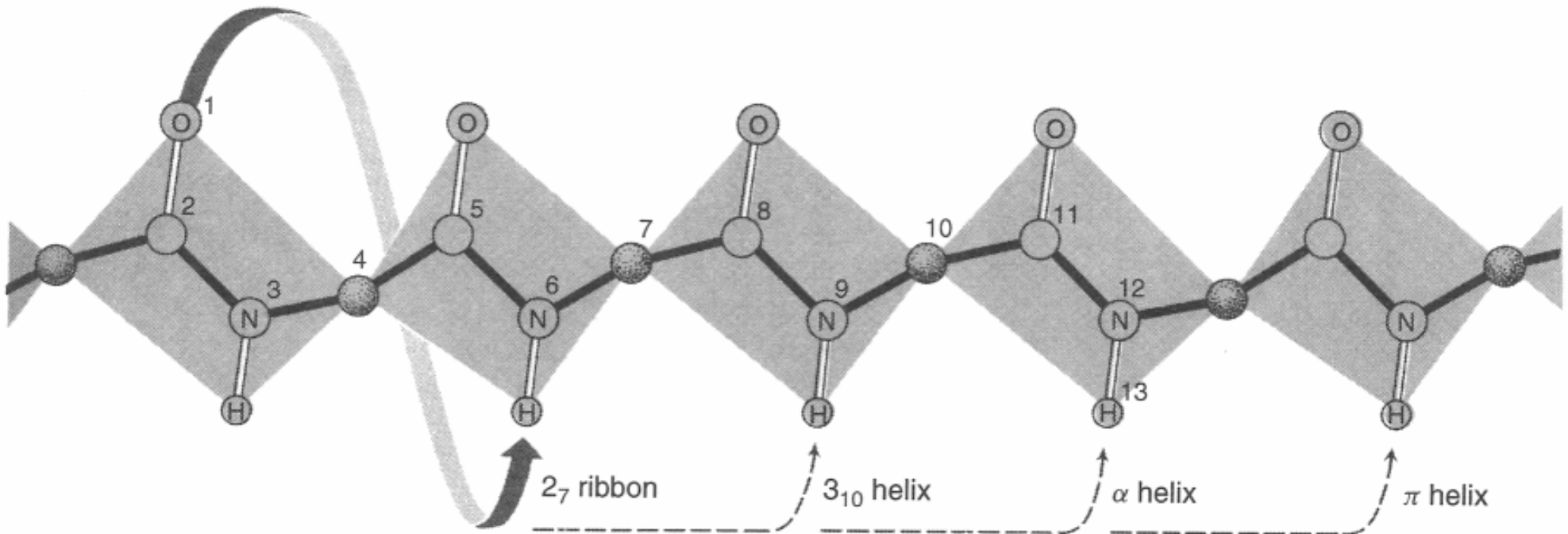
$$E_p = \cos^2(\theta)$$

$$E_t = \begin{cases} [0.9 + 0.1 \sin(2t_i)] \cos(t_o) & 0^\circ < t_i \leq 90^\circ \\ K_1 [K_2 - \cos^2(t_i)] \cos(t_o) & 90^\circ < t_i \leq 110^\circ \\ 0 & 110^\circ \leq t_i \end{cases}$$



CONVENZIONI PER ALFA ELICA

($i \rightarrow i+4$)	alpha elica sito accettore	(>)	O \rightarrow H-N
($i \rightarrow i+3$)	elica 3-10 sito accettore	(>)	O \rightarrow H-N
($i-4 \rightarrow i$)	alpha elica sito donatore	(<)	N-H \rightarrow O
($i-3 \rightarrow i$)	elica 3-10 sito donatore	(<)	N-H \rightarrow O



ESEMPIO DEL PATTERN DEI PONTI-H PER DETERMINARE UNA 4-HELIX OVVERO L'ALFA ELICA

4-turn:

```

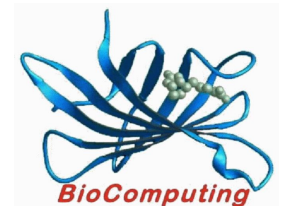
'>'      '4'      '4'      '4'      '<'      notation
-N-C-C--N-C-C--N-C-C--N-C-C--N-C-C      residues
  H   O  N   O  H   O  H   O  H   O
      >-----<      H-bond
  
```

A **minimal helix** is two consecutive N-turns--
for a minimal four helix from residue i to $i+3$:

```

i      <--residue
>444<  and
  >444< overlap to give
>>44<< which defines a helix
HHHH   from  $i$  to  $i+3$ 
  
```

'H' is the notation for a residue in a 4-helix.
Notice that the helix **does not** include the residues
involved in the terminal H-bonds.



Definizione in STRIDE e DSSP di alfa-elica

DSSP

Le definizioni per alfa elica si basano su determinate regole che prevedono una certa regolarità di presenza di ponti idrogeno a distanze regolari e in amminoacidi consecutivi.

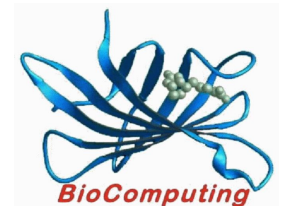
Ad esempio la regola per l'alfa elica è avere un ponte-H tra un AA ed un altro distante da questo 4 AA lungo la catena polipeptidica ed indicato con il simbolo H mentre nell'elica 3_{10} la distanza è di 3 AA ed è indicatoto con G.

Da prove sperimentali si è visto che per avere un'alfa elica occorrono 2 AA consecutivi che formano ponti-H con altrettanti AA a distanza 4 AA ($i \rightarrow i+4$) o ($i-4 \rightarrow i$).

STRIDE

Utilizza le stesse regole di DSSP ma considera anche gli angoli torsionali per vedere la propensione ad un'alfa elica, un beta strand o altro.

Se gli angoli torsionali non sono compatibili con la presenza di un ponte-H che potrebbe far ricadere l'AA in questione in una certa configurazione alfa o beta allora STRIDE non classifica l'AA.



REFERENCE W. KABSCH AND C.SANDER, BIOPOLYMERS 22 (1983) 2577-2637

HEADER ONCOGENE PROTEIN 06-JUN-91 121P

COMPND H-RAS P21 PROTEIN COMPLEX WITH GUANOSINE-5'-[B,G-METHYLENE]

SOURCE HUMAN (HOMO SAPIENS) CELLULAR HARVEY-RAS GENE TRUNCATED AND

AUTHOR U.KRENGEL,K.SCHEFFZEK,A.SCHERER,W.KABSCH,A.WITTINGHOFER,

166 1 0 0 0 TOTAL NUMBER OF RESIDUES, NUMBER OF CHAINS, NUMBER OF SS-BRIDGES(TOTAL,INTRACHAIN,INTERCHAIN)

8891.0 ACCESSIBLE SURFACE OF PROTEIN (ANGSTROM**2)

125 75.3 TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(J) , SAME NUMBER PER 100 RESIDUES

24 14.5 TOTAL NUMBER OF HYDROGEN BONDS IN PARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES

11 6.6 TOTAL NUMBER OF HYDROGEN BONDS IN ANTIPARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES

Angoli torsionali (ϕ, ψ)

#	RESIDUE	AA	STRUCTURE	BP1	BP2	ACC	N-H-->O	O-->H-N	N-H-->O	O-->H-N	TCO	KAPPA	ALPHA	PHI	PSI	X-CA	Y-CA	Z-CA	
1	1	M				120	0, 0.0	2,-0.2	0, 0.0	50,-0.1	0.000	360.0	360.0	360.0	162.6	-5.9	31.9	-6.7	
2	2	T	E	-a	51	0A	61	48,-0.6	50,-2.7	2,-0.0	2,-0.4	-0.425	360.0	-161.0	-62.9	132.1	-4.8	28.9	-4.8
3	3	E	E	-a	52	0A	93	48,-0.2	2,-0.5	-2,-0.2	50,-0.2	-0.926	5.2	-154.9	-114.4	142.4	-4.5	29.7	-1.1
4	4	Y	E	-a	53	0A	13	48,-3.1	50,-2.7	-2,-0.4	2,-0.9	-0.984	7.3	-150.5	-117.8	122.8	-2.5	27.5	1.3
5	5	K	E	-a	54	0A	36	-2,-0.5	71,-2.8	48,-0.2	72,-1.4	-0.818	23.2	-177.6	-97.0	104.1	-3.6	27.6	5.0
6	6	L	E	-ab	55	77A	2	48,-2.4	50,-2.6	-2,-0.9	2,-0.4	-0.807	12.1	-159.7	-105.8	146.7	-0.5	27.0	7.0
7	7	V	E	-ab	56	78A	0	70,-2.0	72,-2.6	-2,-0.3	2,-0.6	-0.989	5.7	-152.9	-130.1	130.4	-0.3	26.7	10.8
8	8	V	E	+ab	57	79A	0	48,-2.6	50,-1.3	-2,-0.4	2,-0.3	-0.917	27.0	167.6	-104.2	120.7	2.9	27.2	12.9
9	9	V	E	+ b	0	80A	0	70,-2.5	72,-2.7	-2,-0.6	2,-0.2	-0.859	11.3	110.3	-132.1	163.3	2.8	25.2	16.2
10	10	G		-	0	0	1	-2,-0.3	72,-0.1	49,-0.3	3,-0.1	-0.769	62.0	-48.8	147.6	166.5	5.3	24.3	18.9
11	11	A	S	> S-	0	0	9	70,-0.5	3,-1.5	78,-0.3	5,-0.3	-0.035	72.5	-71.3	-59.5	161.3	6.2	25.0	22.5
12	12	G	T	3 S+	0	0	56	48,-0.4	-1,-0.2	1,-0.2	77,-0.1	-0.287	113.6	9.3	-60.7	128.1	6.6	28.4	24.1
13	13	G	T	3 S+	0	0	61	-3,-0.1	-1,-0.2	-2,-0.1	-2,-0.1	0.488	83.8	121.2	85.1	7.0	9.6	30.4	23.1
14	14	V	S	< S-	0	0	3	-3,-1.5	70,-0.1	67,-0.1	-2,-0.1	0.656	88.2	-99.1	-77.9	-14.8	10.9	28.2	20.2
15	15	G	S	> S+	0	0	15	-4,-0.2	4,-2.6	66,-0.1	5,-0.2	0.637	71.8	144.9	108.4	24.2	10.6	31.0	17.7
16	16	K	H	> S+	0	0	12	-5,-0.3	4,-2.1	1,-0.2	5,-0.1	0.933	81.4	41.1	-53.7	-50.0	7.3	30.4	15.9
17	17	S	H	> S+	0	0	26	2,-0.2	4,-2.9	1,-0.2	5,-0.3	0.902	112.1	53.2	-68.2	-44.3	6.7	34.1	15.6
18	18	A	H	> S+	0	0	11	1,-0.2	4,-2.0	2,-0.2	-1,-0.2	0.893	109.8	50.4	-61.1	-37.2	10.2	35.1	14.7
19	19	L	H	X S+	0	0	1	-4,-2.6	4,-2.3	2,-0.2	-2,-0.2	0.969	112.7	45.5	-62.7	-52.2	10.2	32.5	11.9
20	20	T	H	X S+	0	0	0	-4,-2.1	4,-3.2	-5,-0.2	5,-0.3	0.898	113.5	48.0	-60.1	-41.3	6.9	33.8	10.5

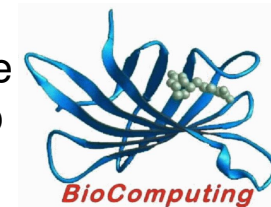
Accessibilità

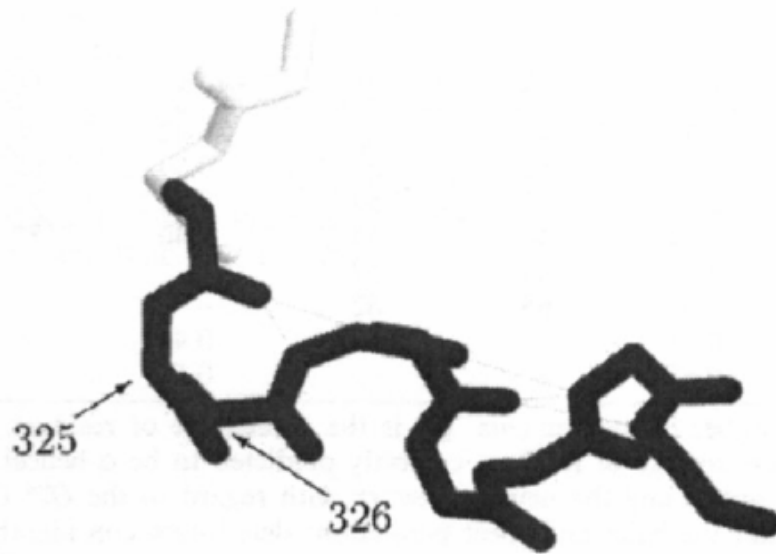
Indica >,<,X accettore o donatore sulla base delle energie N-H-O

Struttura secondaria

Numerazione residui

Le due migliori E dei ponti-H con la backbone. 2 per accettore e 2 per i donatori; se una tra tutte prevale allora si decide se è ">" o "<" oppure "X" se paragonabili. Il primo numero indica l'AA con cui fa ponte, il secondo l'energia.

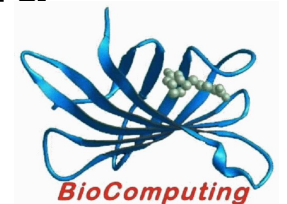




322	F	<
323	K	>
324	S	H >
325	K	H 4
326	D	H 4
327	S	H X
328	V	H >X
329	P	H 3>
330	K	H 3X
331	L	H <X
332	V	H X

Fig. 3. This example shows two amino acids (#325 and #326), which have been assigned 'H' by DSSP, but without hydrogen bonds in the helix. On the left is a view of the backbone structure and hydrogen bonds, and on the right a cut-out of the corresponding DSSP assignments (> indicates a $i \rightarrow i+(3 \text{ or } 4)$ hydrogen bond, < indicates an $i-(3 \text{ or } 4) \leftarrow i$ hydrogen bond and X indicates both). The protein shown has the PDB name 8adh.

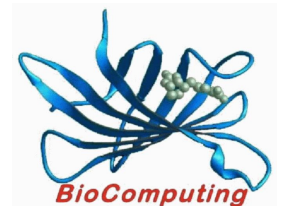
- Le strutture secondarie assegnate da *DSSP* ed altri metodi (p.es. *STRIDE*) sono identiche solo nel **85-90% dei casi**.
- Questo è quindi il limite **teorico** per qualsiasi metodo di predizione.
- *DSSP* è comunque il “*gold standard*” per determinare la struttura secondaria di strutture PDB.



Accuratezza

- La misura più intuitiva e diffusa è il Q_3 , ossia la percentuale di residui correttamente predetta.
- La formula è:
$$Q_3 = 100 * 1/N * \sum_{i=\alpha,\beta,loop} M_i$$

N è il numero totale di residui,
 M_i sono le predizioni corrette (α , β , $loop$).
- Un'altra misura utilizzata è il SOV (*segment overlap*) che tende a penalizzare ulteriormente la presenza e/o assenza di interi elementi di struttura secondaria. (La formula è troppo complessa per essere spiegata brevemente)
- Generalmente i valori di SOV sono ca. 5-6% sotto quelli di Q_3 .

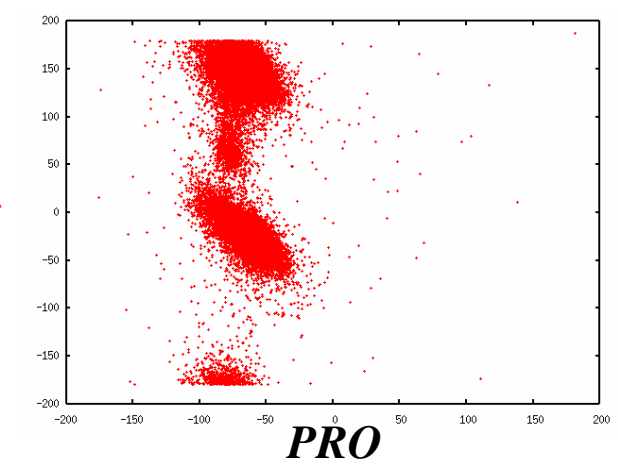
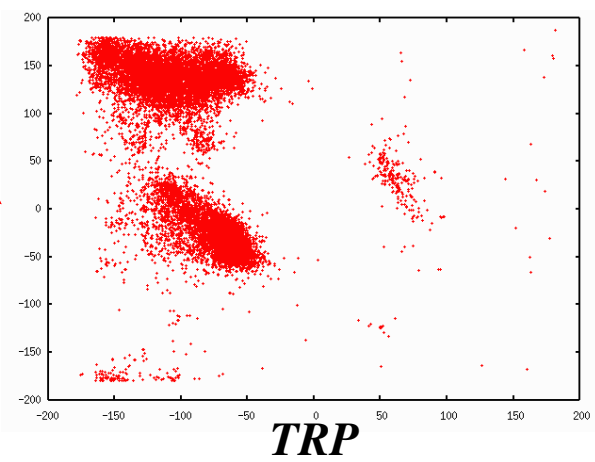
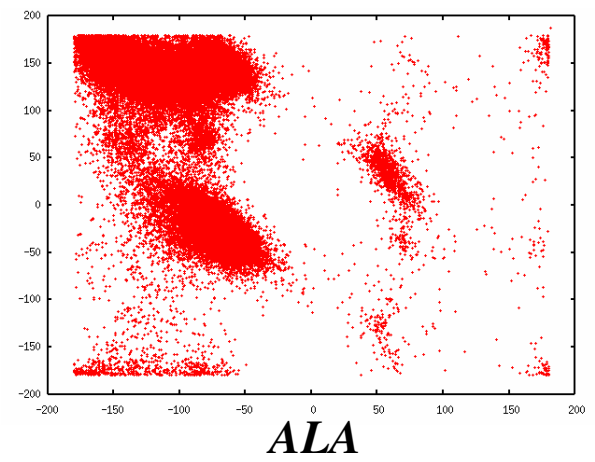
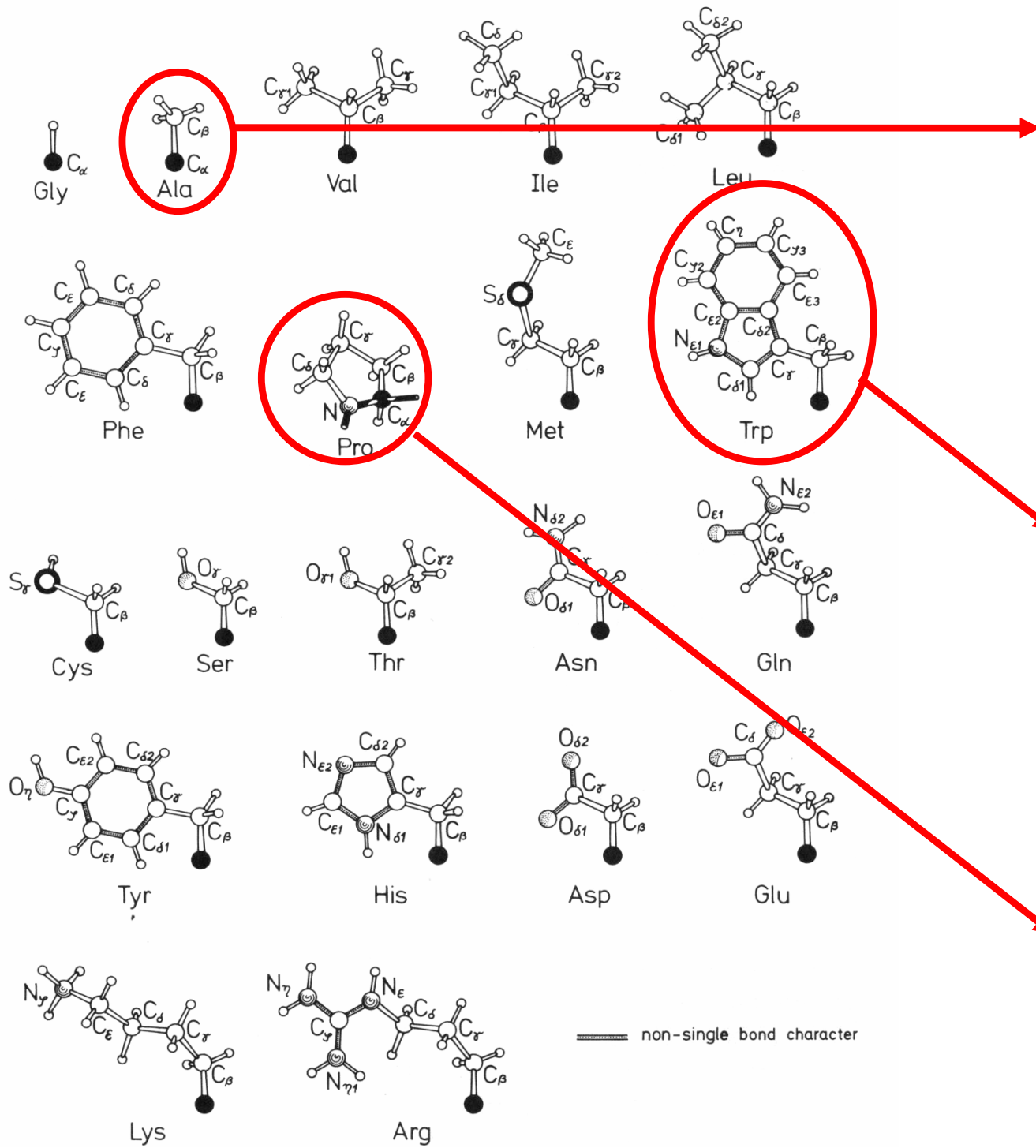




METODI PREDITTIVI DI STRUTTURA SECONDARIA

Sono metodi sviluppati per la predizione della struttura secondaria a partire solo dalla sequenza amminoacidica e quindi in assenza di strutture PDB associate alla sequenza stessa.

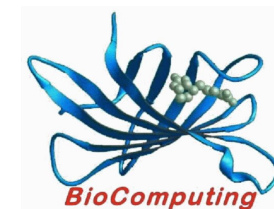
BioComputing



Metodo di Chou & Fasman (1974)

Gli aminoacidi hanno propensioni diverse a formare strutture α -eliche e filamenti β . La prolina p.es. interrompe le α -eliche. Regole predittive che utilizzano parametri di propensione conformazionale ricavati da un'analisi statistica dei dati cristallografici. (15 proteine, 2473 residui)

	Blout <i>et al.</i> , 1960 (328)	Kotelchuck and Scheraga, 1968 (363)	Lewis <i>et al.</i> , 1970 (368)	Robson and Pain, 1971 (346)	Chou and Fasman, 1974 (340)	Finkelstein and Ptitsyn, 1976 (371)
A Ala	(H)	H	I	+0.09	1.45	1.08
C Cys	C	H	I	+0.03	0.77	0.95
D Asp	H	C	B	-0.02	0.98	0.85
E Glu	H	H	H	+0.12	1.53	1.15
F Phe	(H)	H	H	+0.03	1.12	1.10
G Gly	—	Indifferent	B	-0.05	0.53	0.55
H His	(H)	H	I	+0.08	1.24	1.00
I Ile	(C)	H	H	+0.07	1.00	1.05
K Lys	(H)	C	I	-0.03	1.07	1.15
L Leu	H	H	H	-0.11	1.34	1.25
M Met	H	H	H	-0.10	1.20	1.15
N Asn	(C)	C	I	-0.04	0.73	0.85
P Pro	—	Special	B	—	0.59	—
Q Gln	(H)	H	I	+0.07	1.17	0.95
R Arg	(H)	H	I	+0.02	0.79	1.05
S Ser	C	C	B	-0.07	0.79	0.75
T Thr	(C)	H	I	-0.01	0.82	0.75
V Val	C	H	I	+0.04	1.14	0.95
W Trp	(H)	C	H	+0.10	1.14	1.10
Y Tyr	(H)	C	H	-0.02	0.61	1.10

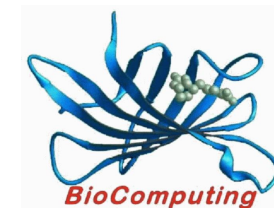


Metodo di Chou & Fasman (1974)

α ELICA			FOGLIETTO β		
favorenti					
aa	P_{α}	H α	H β	aa	P_{β}
Glu Ala Leu	1.53 1.45 1.34			Met Val Ile	1.67 1.65 1.60
deboli favorenti					
His Met Gln Trp Val Phe	1.24 1.20 1.14 1.14 1.14 1.12	h α	h β	Cys Tyr Phe Trp Gln Leu Thr	1.30 1.29 1.28 1.26 1.23 1.22 1.20
indifferenti favorenti					
Lys Ile	1.07 1.00	l α	l β	Ala Arg	0.97 0.90
indifferenti destruenti					
Asp Thr Ser Arg Cys	0.98 0.82 0.79 0.79 0.77	i α	i β	Gly Asp	0.81 0.80
deboli destruenti					
Asn Tyr	0.73 0.61	b α	b β	Lys Ser His Asn Pro	0.74 0.72 0.71 0.65 0.62
destruenti					
Pro Gly	0.59 0.53	B α	B β	Glu	0.26

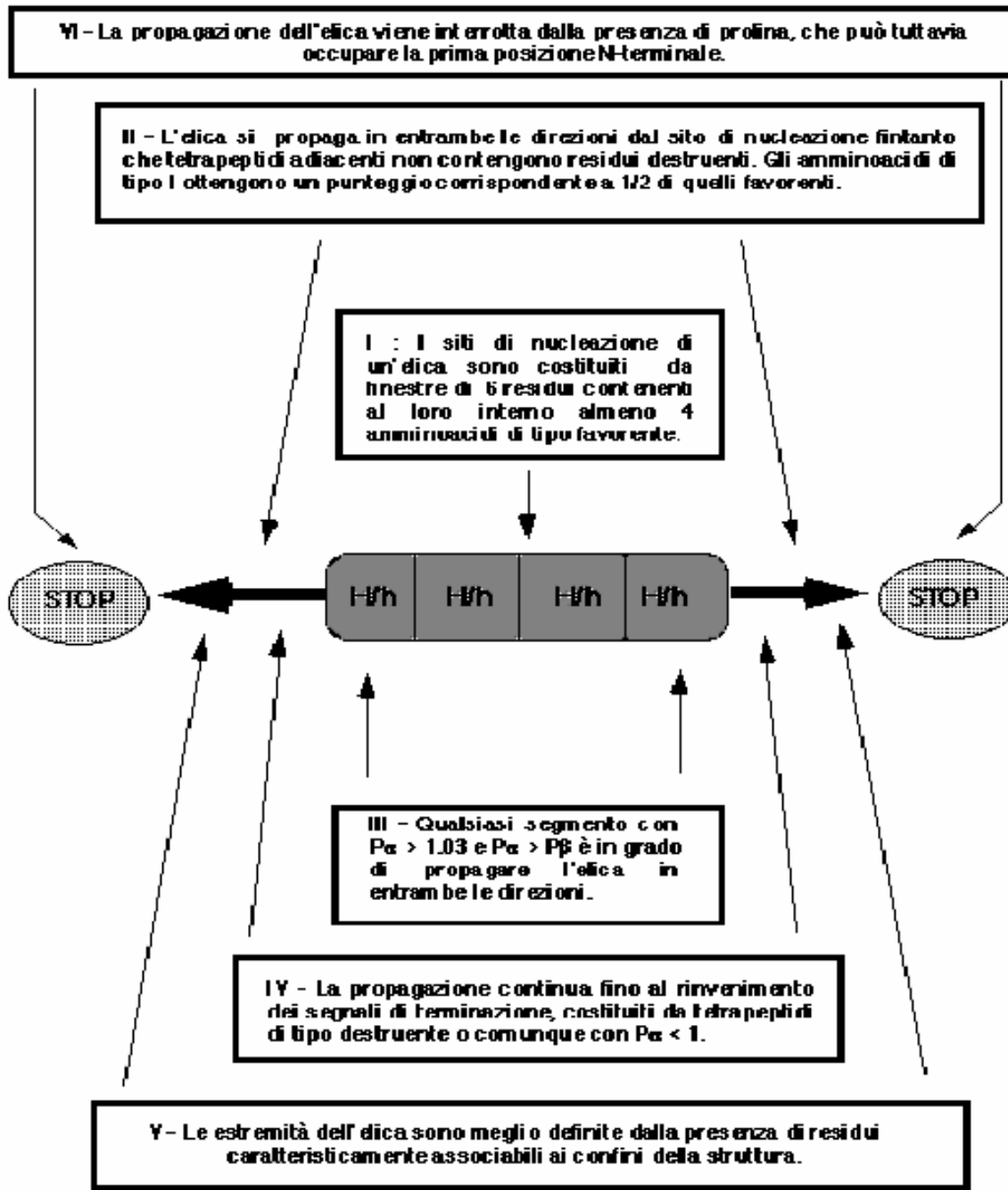
Se N_j è il numero totale di residui nello stato j (alfa-elica, foglietto beta, coil) N_T il numero totale di residui, n_{ij} la frequenza di un dato aminoacido i a trovarsi nello stato j e n_i la frequenza dell' i -esimo aminoacido nel campione, si può definire la propensione del residuo i -esimo a comparire nella struttura secondaria di tipo j come (P_{ij}):

$$P_{ij} = (n_{ij}/n_i) / (N_j/N_T)$$



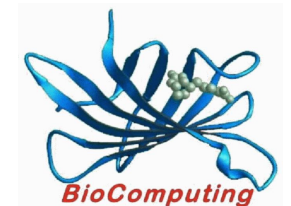
TEORIA DI FORMAZIONE DELLE STRUTTURE SECONDARIE

CHOU E FASMAN - TEORIA DI FORMAZIONE DELL' ELICA



ESEMPIO DELL'ALFA ELICA

Il concetto di base è che su frammenti più o meno estesi (*sliding windows*) si calcola la propensione media alla struttura alfa o beta secondo un valore soglia e si attribuisce a quella regione la propensione.



ALTRI METODI BASATI SULL'ANALISI DELLA SINGOLA SEQUENZA

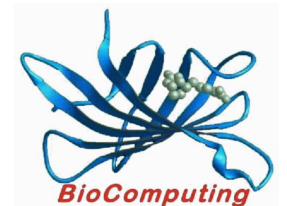
Metodi di I° GENERAZIONE

In realtà questi metodi calcolano la propensione alla struttura secondaria basandosi su singola sequenza e con un contesto locale molto ridotto e si attestano intorno al 50-60% (anni '80).

- *Il metodo GOR (Garnier-Osguthorpe-Robson, 1978) è una modificazione del metodo precedente e valuta finestre di contesto locale più ampie (Q_3 50-60%).*

Metodi di II° GENERAZIONE

In auge fino agli anni 90 questi metodi prendono in considerazione contesti locali più ampi (fino a 51 AA) ed anche allineamenti multipli sebbene anche questi metodi soffrono di alcuni inconvenienti e si attestano poco oltre il 60% di Q_3 .



QUALI SONO I PRINCIPALI PROBLEMI ?

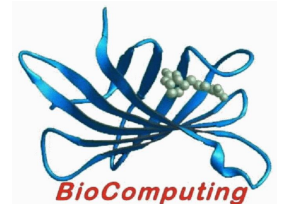
Si è osservato che il contesto locale soltanto contribuisce per il 65% alla formazione della struttura secondaria:

- a) I segmenti di SS erano in media più corti rispetto alla realtà;
- b) I foglietti beta predetti in modo inefficiente (predizione casuale) perchè i ponti-H sono tra AA distanti in sequenza.

SOLUZIONE PROPOSTA: Metodi di III° Generazione

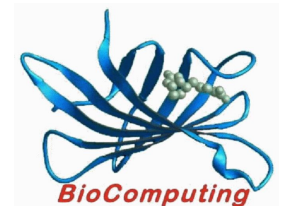
Numero elevato di allineamenti multipli e implementazione di reti neurali che cercano di “imparare” la situazione dalle seguenti osservazioni:

1. La SS può rimanere invariata sostituendo anche gran parte degli AA. Ma è anche vero che poche sostituzioni la possono cambiare (prolina rompe l'alfa elica);
2. Questo si riflette nell'allineamento multiplo in cui è possibile distinguere “le mutazioni neutrali” dalle posizioni “chiave” che devono essere mantenute.



SOLUZIONE PROPOSTA ***METODI DI III° GENERAZIONE***

Le reti neurali sembrano essere la risposta migliore perché sono in grado di imparare da un set di dati composto da allineamenti multipli e ricercano relazioni tra posizioni distanti all'interno dell'allineamento multiplo che sono fondamentali per la formazione della struttura secondaria. Le informazioni “evolutive” dell'allineamento multiplo hanno dato uno dei maggiori contributi all'aumento della accuratezza di tali metodi.



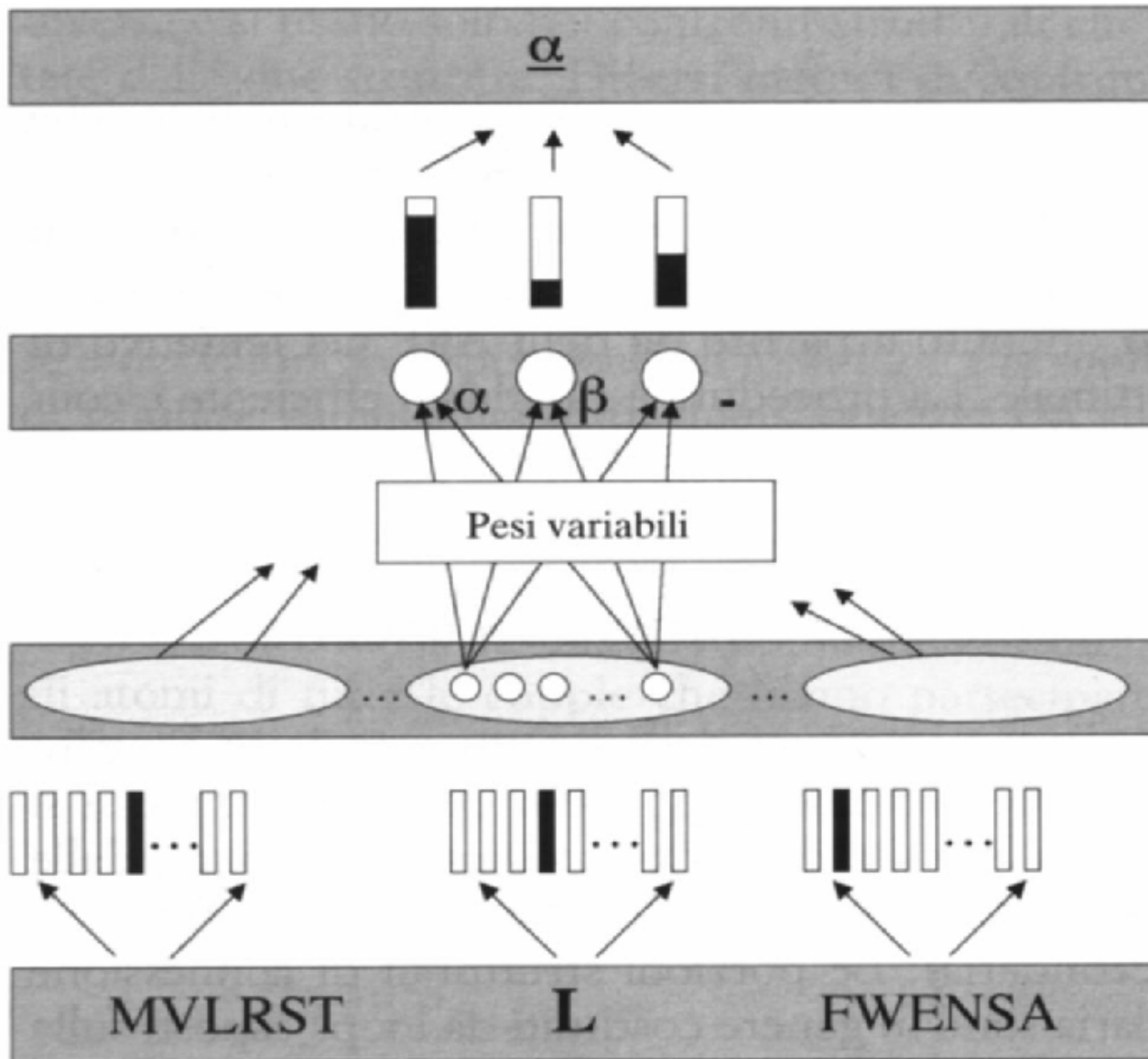


***I risultati migliorano notevolmente
utilizzando
metodi di machine learning.***

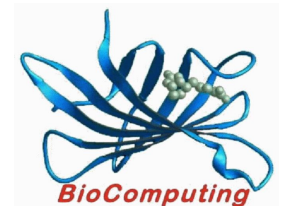
(allineamenti multipli, reti neurali ecc.)

BioComputing

Reti neurali



- Tutti i migliori metodi di predizione di struttura secondaria (eccetto i metodi *consensus*) utilizzano reti neurali.
- La parametrizzazione delle reti neurali richiede molti esempi (**fino a 2000**) di proteine *non omologhe*.
- Per la predizione del residuo i della proteina si utilizza il contesto locale (p.es. $i-6, \dots, i-1, i, i+1, \dots, i+6$)
- Ogni residuo è codificato in modo sparso. 21 unità per ogni posizione: 20 per ogni tipo di residuo, uno per l'assenza (*gap*).



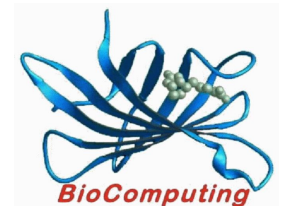
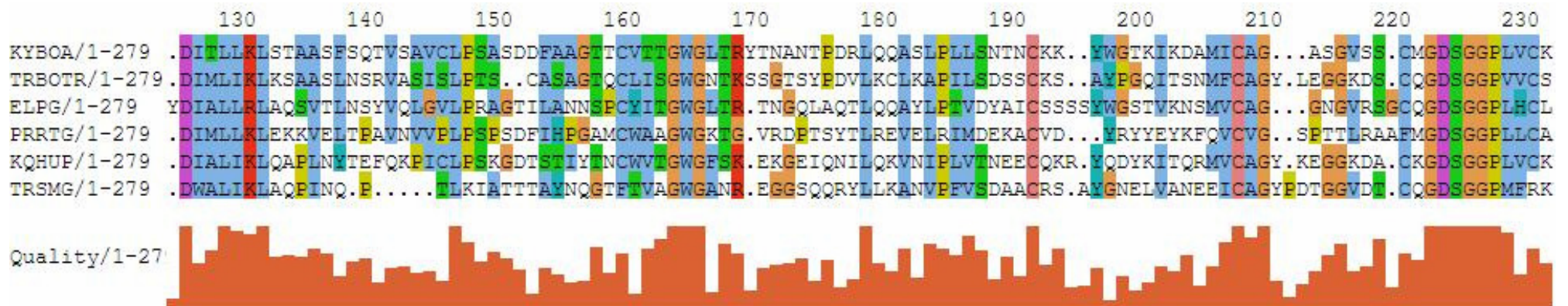
Profile HeiDelberg (PHD)

Il primo metodo di “terza generazione”. (Rost & Schneider, 1993)

- Q_3 al 72% ca.

Due novità importanti:

- Utilizzo di informazioni sulle sequenze omologhe (estratte da HSSP).
- Utilizzo di tre livelli di predizione per filtrare le predizioni e ridurre gli errori locali.

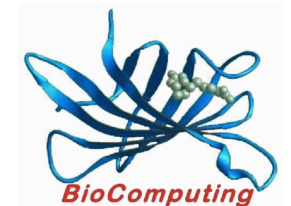
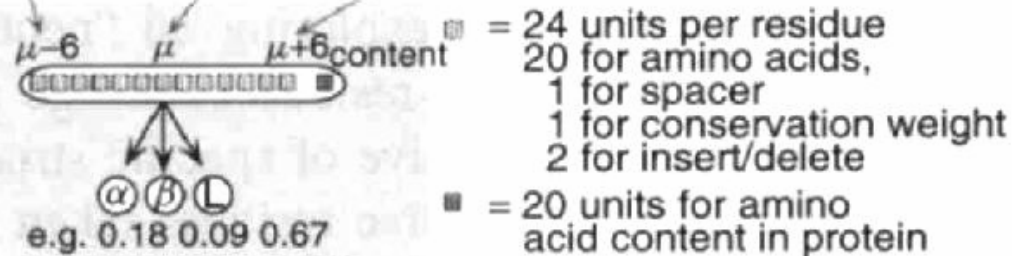


PHD

Utilizza in input un profilo proveniente da un allineamento multiplo della sequenza di cui si vuole determinare la struttura secondaria con le proteine con cui condivide zone di similarità e ricavate da una ricerca in banca dati con BLAST. In alternativa può accettare un allineamento multiplo già fatto e costruito ad hoc dall'utente.

INPE	L	L	L	L	L	E	E	E	E	E	E	E	E	E	E	E	E	H	H	H			
SH	N	S	T	N	K	D	W	W	K	V	E	V	N	D	R	Q	G	F	V	P	A	A	Y
a1	N	K	S	N	P	D	W	W	E	G	E	L	N	G	Q	R	G	V	F	P	A	S	Y
a2	E	E	H	.	G	E	W	W	K	A	K	s	s	K	R	E	G	F	I	P	S	N	Y
a3	R	S	T	.	G	D	W	W	L	A	r	v	T	G	R	E	G	Y	V	P	S	N	Y
a4	F	S	F	F	G	V	e	v	D	D	L	Q	V	F	V	P	P	A	Y
V	0	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	20	20	60	0	0	0	0
L	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W	0	0	0	0	0	0	80	80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0	20	20	0	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	60	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R	20	0	0	0	0	0	0	0	0	0	0	0	0	0	60	20	0	0	0	0	0	0	0
K	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	0	0	0
E	20	20	0	0	0	25	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0
N	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	0	0	75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ndel	0	0	1	3	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Nins	0	0	0	0	0	0	0	0	0	0	0	2	3	1	0	0	0	0	0	0	0	0	0

First level:
sequence-to-structure



PHD

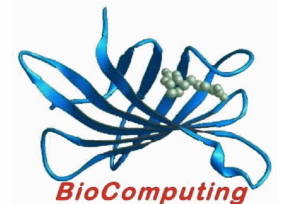
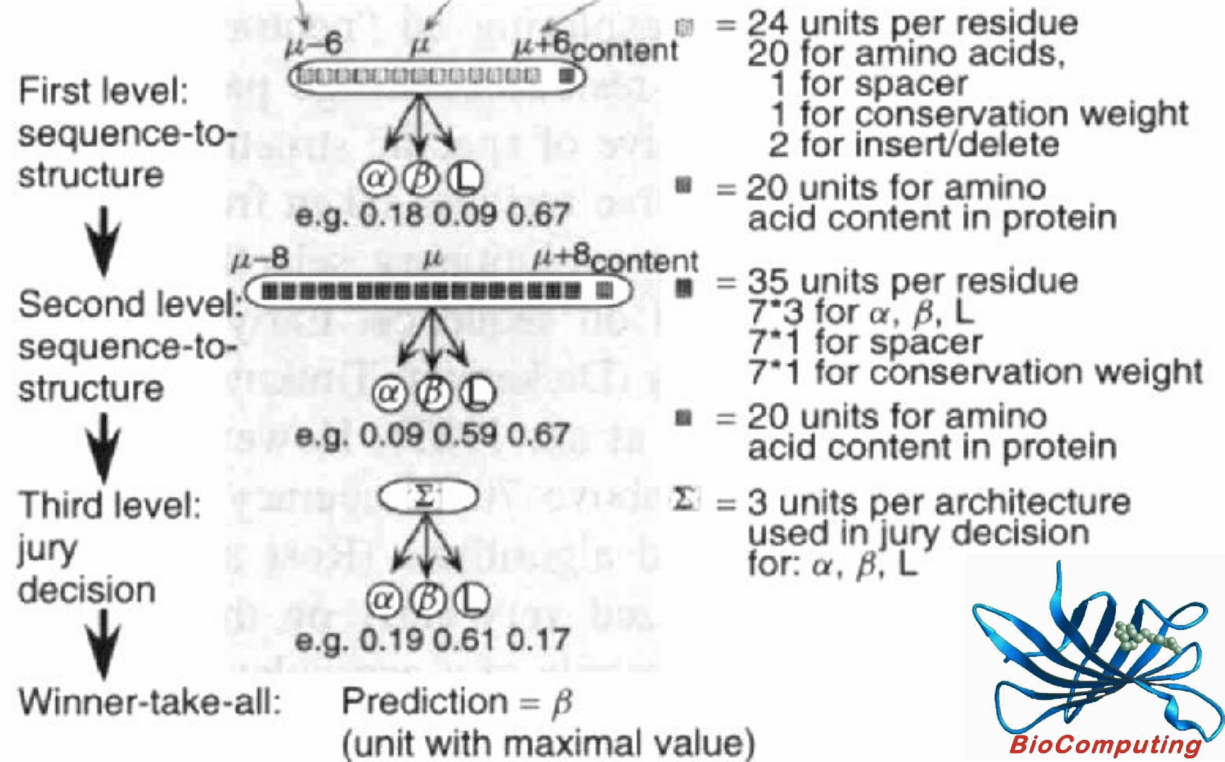
La novità sta nell'utilizzare più di una rete neurale: In questo modo si possono eliminare piccoli errori locali (p.es. α -eliche di lunghezza inferiore a 4 residui).

INSPE	L	L	L	L	L	E	E	E	E	E	E	E	E	E	E	E	E	E	H	H	H			
SHN	S	T	N	K	D	W	W	K	V	E	V	N	D	R	Q	G	F	V	P	A	A	Y		
a1	N	K	S	N	P	D	W	W	E	G	E	L	N	G	Q	R	G	V	F	P	A	S	Y	
a2	E	E	H	.	G	E	W	W	K	A	K	s	s	K	R	R	E	G	F	I	P	S	N	Y
a3	R	S	T	.	G	D	W	W	L	A	r	v	T	G	R	E	G	Y	V	V	P	S	N	Y
a4	F	S	.	.	.	F	F	G	V	e	v	D	D	L	Q	V	F	V	P	P	A	Y		
V	0	0	0	0	0	0	0	0	0	40	0	60	0	0	0	0	20	20	60	0	0	0	0	0
L	0	0	0	0	0	0	0	0	20	40	0	20	0	0	0	20	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	20	0	0	0	0	0	0	20	20	0	0	0	0	0	0	0	0	60	20	0	0	0	0	20
W	0	0	0	0	0	0	80	80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	80
G	0	0	0	0	0	0	0	0	20	20	0	0	0	40	0	0	80	0	0	0	0	0	0	0
A	0	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	100	20	40	40	0
P	0	0	0	0	0	0	0	0	0	0	0	20	20	0	0	0	0	0	0	0	0	0	0	0
S	0	0	60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	0	20	0	0	0	0	0	0	40	0	0	0	0	20	20	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	20	20	0	0	0	0	0	0	0	0	60	0	0	0	20	40	0	0	0	0	0	0	0	0
N	40	0	0	0	100	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	40	0
D	0	0	0	0	0	0	70	0	0	0	0	0	20	40	0	0	0	0	0	0	0	0	0	0
Nhel	0	0	1	3	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Nhrs	0	0	0	0	0	0	0	0	0	0	2	3	1	0	0	0	0	0	0	0	0	0	0	0

CCCC**HH**CC**CH**HE**HHH**



CCCC**CC**CC**CH**HH**HHH**



Output di PHD

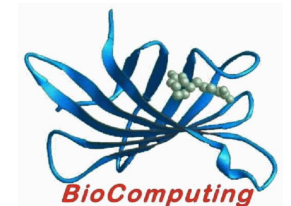
Predizione di struttura secondaria per la proteina ras umana.

Nella prima riga dell'output è riportata la numerazione relativa alla sequenza di input. Nella riga successiva (*AA*) è riportata la sequenza di riferimento (in questo caso la proteina ras). Nella terza riga dell'output (*PHD sec*) si trova la predizione della struttura secondaria (*E* indica la struttura extended, ovvero beta strand, *H* indica helix, ovvero alpha-elica). La riga successiva (*Rel sec*) offre i valori dell'affidabilità da 0 (bassa) a 9 (alta) della predizione, posizione per posizione. Nelle righe sottostanti viene riportato il dettaglio della predizione, cioè le preferenze per ognuno dei tre possibili stati (beta; alpha; o loop).

```

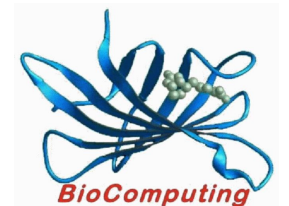
      .....1.....2.....3.....4.....5.....6
AA      |MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAG|
PHD sec | EEEEEEEE      HHHHHHHHHHH      EEEEE      HHHHHHH      |
Rel sec | 938999997389866589987789761323468771112352599447744156773166|
detail:
prH sec | 000000000000112689987888874211210014443211100100122367885522|
prE sec | 068999998310000000011100101232110100011115688621101210000000|
prL sec | 931000001688877210000000014545578774444563200268766311113477|
subset: SUB sec | L.EEEEEEE.LLLLLHHHHHHHHHH.....LLLL.....L.EEE..LL...HHH..LL|

      .....7.....8.....9.....10.....11.....12
AA      |QEEYSAMRDQYMRTGEGFLCVFAINNTKSFEDIHQYREQIKRVKDSDDVPMVLVGNKCDL|
PHD sec | HHHHHHHHHHHHH      EEEEE      HHHHHHHHHHHHHHH      EEEEE      HH|
Rel sec | 225799999998713662699997299964789999999998634699828987357125|
detail:
prH sec | 4568999999987531010000000002688999999999875210000000001456|
prE sec | 011000000000000113789998400000000000000000111000058888521000|
prL sec | 531100000001145675100001599872110000000001126799831001377442|
subset: SUB sec | ..HHHHHHHHHH..LL.EEEEE.LLLL.HHHHHHHHHHH..LLLL.EEEE.LL..H|
```



Oltre PHD

- Negli ultimi anni sono usciti nuovi programmi in grado di incrementare il valore medio di Q_3 fino al 76-77% ca.
- *JPRED* (Cuff & Barton, 1999) è un esempio di metodo *consensus*. Invece di creare un nuovo predittore, si cerca di combinare i risultati di altri metodi di successo per migliorare il risultato finale. Non è più stato aggiornato da oltre tre anni.
- Una evoluzione di PHD è PROFsec (utilizza profili di PSI-BLAST).



PSIPRED *(Jones, 2000)*

Raw profile from PSI-BLAST Log File

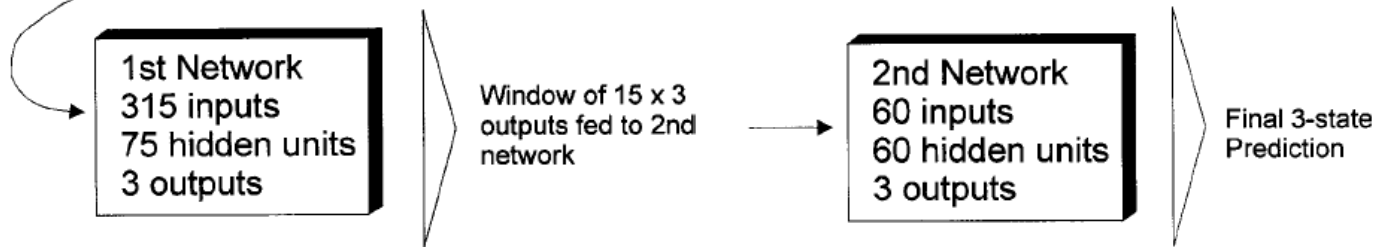
Position-based scoring matrix used

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
-3	-4	-4	-4	-3	-4	-4	-4	-2	-1	-1	-4	-1	8	-5	-3	-3	0	2	-2
0	-1	-1	3	-4	3	4	1	-1	-4	-4	0	-3	-4	-2	-1	-2	-4	-3	-3
0	-1	2	1	-3	4	0	-1	-2	-4	-3	1	-2	-4	-2	2	0	-4	-3	-3
-2	-3	-4	-5	-2	-3	-4	-6	-4	0	6	0	0	-1	-4	-3	-2	-4	-2	0
0	-3	-1	-2	-3	0	-2	4	-3	-3	0	-2	-2	-4	-3	3	1	-4	-4	-3
0	2	0	4	-4	1	2	1	-2	-4	-4	0	-3	-4	-3	1	-2	-5	-4	-4
-1	5	3	-2	-4	-1	-1	1	-2	-1	-4	1	-3	-4	-3	1	-2	-5	-4	-4
-2	-3	-4	-5	-3	-3	-4	-5	-4	3	4	-1	1	2	-4	-3	-2	-3	-1	0
-2	3	2	-2	-4	2	1	-3	-2	-3	-3	1	1	-4	-3	2	1	-4	-3	-1
0	2	3	1	-4	0	0	0	-2	-4	-4	1	-3	-4	-3	2	0	-5	-4	-4
5	-3	-3	-3	-2	-3	-3	-2	-3	1	-2	-3	-2	1	-3	0	1	-4	-2	0
-1	-4	-5	-5	-3	-4	-4	-5	-4	3	3	-4	2	3	-5	-3	-2	5	-1	2
0	3	3	0	-4	3	0	1	-2	-4	-4	1	-3	-4	-3	1	-1	-4	-3	-4
-1	0	1	0	-4	1	-1	-1	-2	-4	-3	5	-2	0	-3	0	-2	-4	0	-3
-2	-3	-1	-5	-3	-3	-4	-5	-4	3	4	0	4	2	-4	-3	-2	-3	-2	0
0	3	0	-2	-3	-1	0	0	-2	0	0	1	0	-1	-3	2	0	-4	-3	0
-1	1	3	-2	-4	0	-2	4	-2	-4	-4	0	-3	0	-3	0	0	-3	0	-4

Window of 15 rows

A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	
0.4	0.3	0.3	0.3	0.2	0.9	0.3	0.3	0.4	0.4	0.4	0.3	0.4	0.9	0.1	0.4	0.4	0.5	0.7	0.4	
0.3	0.2	0.3	0.8	0.4	0.3	0.7	0.1	0.6	0.2	0.4	0.3	0.5	0.2	0.1	0.4	0.8	0.2	0.3	0.2	
0.1	0.1	0.4	0.3	0.5	0.1	0.1	0.3	0.1	0.1	0.4	0.2	0.4	0.9	0.3	0.4	0.4	0.9	0.3	0.6	
0.6	0.3	0.3	0.1	0.3	0.5	0.5	0.2	0.1	0.4	0.4	0.3	0.6	0.9	0.1	0.5	0.1	0.5	0.7	0.4	
.

15 x 20 scaled inputs to 1st network



Utilizza i profili di PSI-BLAST ed il training è avvenuto su più proteine rispetto a PHD

Utilizza una finestra a 15 amminoacidi (nella versione originale) scelta dopo vari training a finestre diverse.

Usa due reti neurali. La seconda decide la propensione per struttura secondaria del residuo sulla base dell'input della prima.

Predizione intorno ad oltre il 75% corrette.

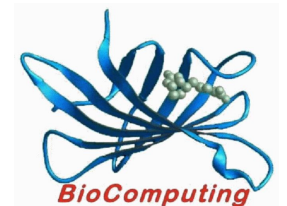


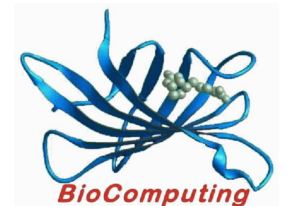
Figure 1. An outline of the PSIPRED method, which shows how the PSI-BLAST score matrices are processed.

PROTOCOLLO DI PSIPRED

In order to maximise the effectiveness PSI-BLAST in producing very sensitive profiles, custom sequence data bank was constructed for the present application.

Firstly, a large non-redundant protein sequence data bank was compiled extracting non-identical sequences from a number of publicly available data banks. This databank is then filtered to remove regions with very low information content. A custom program is used to further filter the data bank order to remove transmembrane segments, and regions which are likely to form coiled-coil structures.

1. The final position-specific scoring matrix (log-odds values) from PSI-BLAST (after three iterations with BLOSUM62) is used as input to the neural network.
2. A window of 15 amino acid residues was found to be optimal, and thus the final input layer comprises 315 input units, divided into 15 groups of 21 units. The extra unit per amino acid is used to indicate where the window spans either the N or C terminus of the protein chain.
A large hidden layer of 75 units was used, 3 units making the output layer where the units represent the three-states of secondary structure (helix, strand or coil).
3. A second network is used to filter successive outputs from the main network.
A sort of 3D secondary structure Jury that judges the prediction of each single aminoacid based on its first result of alpha-beta-coil propensity derived from the first Network. As only three possible inputs are necessary for each amino acid position, this network has an input layer comprising just 60 input units, divided into 15 groups of four. Again the extra input in each group is used to indicate that the window spans a chain terminus. For this network, a smaller hidden layer of 60 units was used. 3 units making the output layer where the units represent the three-states of secondary structure (helix, strand or coil).



ULTERIORI SVILUPPI - ssPRO

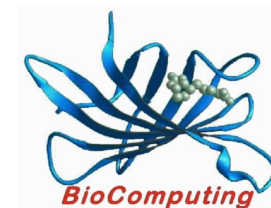
Uno dei limiti dei metodi come PHD o PSIPRED è che sono limitati dalla lunghezza fissa dell'intorno locale all'AA da considerare per la propensione ad una certa struttura secondaria. *Difficile predizione dei foglietti beta.*

PSI-PRED ha seguito la strada di aumentare le informazioni basandosi sui profili di PSI-BLAST ma utilizza sempre una rete feed-forward.

Per migliorare questo aspetto si è pensato di utilizzare la proteina in toto con un nuovo tipo di algoritmo basato sulle reti neurali bidirezionali ricorrenti.

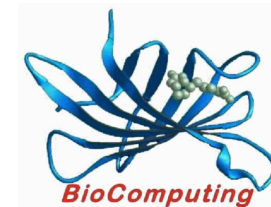
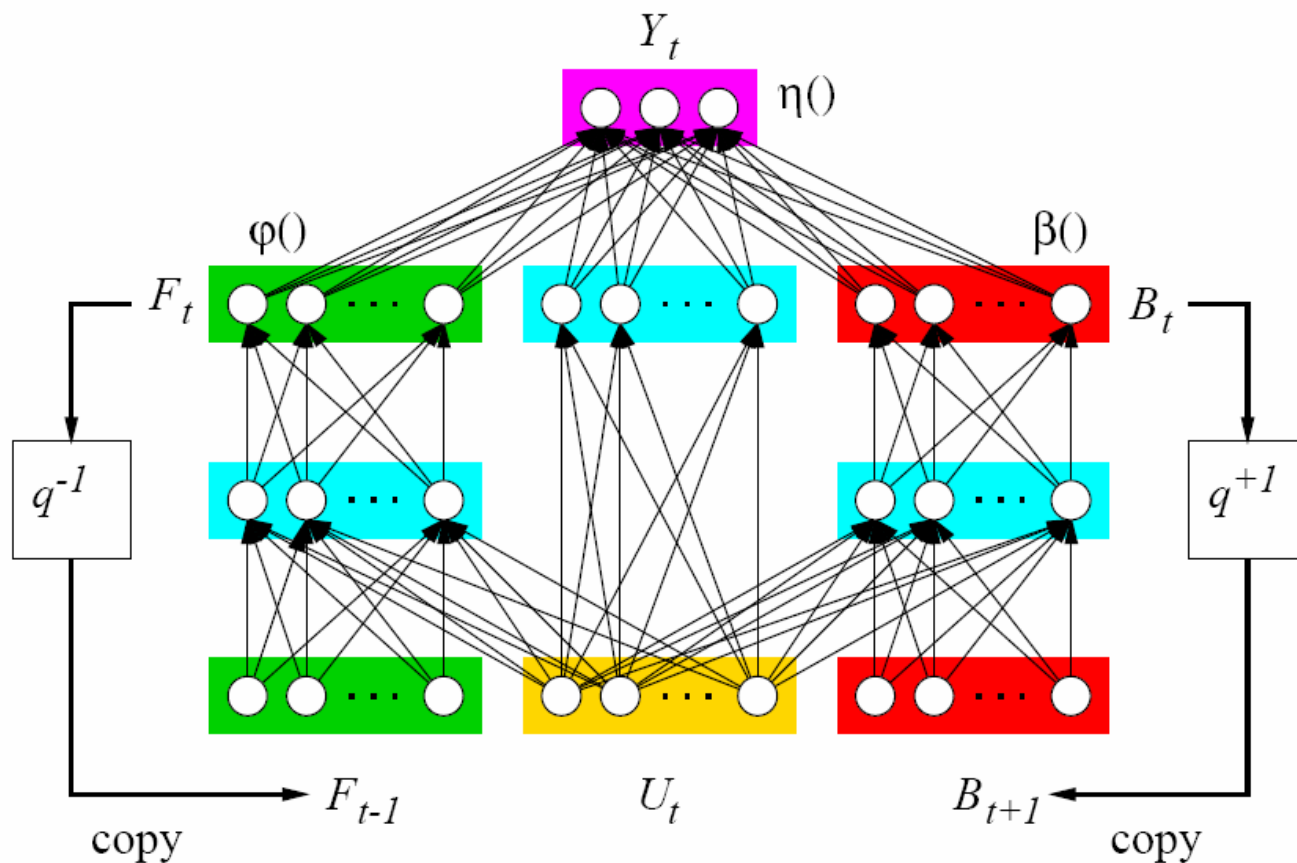
BRNN - Bidirectional Recurrent Neural Network

(Baldi et al. 1999, Pollastri et al. 2001)



ULTERIORI SVILUPPI - ssPRO

La novità sta nel fatto che la finestra dell'intorno locale viene dinamicamente allargata nel processo di valutazione alla propensione per una certa struttura secondaria. Applicata al profilo di PSI-BLAST è in grado di cogliere significative relazioni distanti per strutture, ad esempio, beta-strand.



Differenze tra i metodi

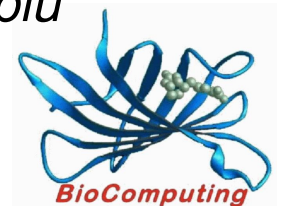
seq	SAFTVWSGPGCNNRAERYSKCGCSAIHQKGGYDFSYTGQTAALYNQAGCSGVAHTRFGSSARACNPFGWKSIFIQC
dssp	-EEEE-----EEE---EEEE-----EEEE-----EEEE--HHH---EEEE-----E-----EEEE--
jpred	---EEE-----EE-----EEE-----EE-----EEEE--
phd	-EEEE-----HHHH-----EEEE-----EE--HHHHHHH-----EEEE-----EEEEEE--
phdpsi	-EEEE-----HHHH-----EEEE-----EE--HHHHHHH-----EEEE-----EEEEEE--
prof_king	-EEEE-----EEEE-----EEE---EEEE-----EEE-----H-----EEEE--
profsec	-EEEE-----HHHEEE-----EEEE-----EE--HHHHHHHH-----EEEEEE-----EEEEEE--
psipred	-EEEE-----HHHHHHHHHHHHHH-----HHHH-----EEEE--
pssp	-EEEEEE-----HH-----EE-----EEEE-----HHH-----EEEE-----EEEE--
samt99_sec	--EEEE-----HHHH-----EEE-----EEE---HHHHHH-----EE-----EEEEEE--
sspro1	--EEEE-----HHHHHH-----HHHH-----EEE-----EEEEEE--

PHD è stato superato per due motivi principali:

- l'utilizzo di sequenze sempre più remote
- l'allenamento effettuato con un numero molto maggiore proteine non omologhe (= *crescita dei database*).

PSIPRED è in grado di predire meglio le α -eliche, mentre **ssPRO** funziona meglio con i β -strand.

Un'idea abbastanza ovvia sarebbe quindi quella di combinare i risultati di più metodi per ottenere predizioni più affidabili.



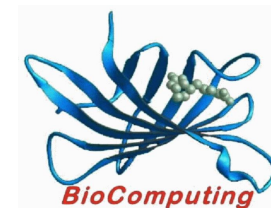
I METODI CONSENSO

Method	Q_3		Q_H		Q_E		Q_L		SOV	
	μ	err_σ	μ	err_σ	μ	err_σ	μ	err_σ	μ	err_σ
(a)										
PSIPRED (PS)	74.24	1.08	76.34	2.20	69.61	2.82	74.73	1.23	70.23	1.63
SAM-T99 (SA)	73.97	0.96	77.02	2.29	69.58	2.43	71.84	1.55	67.67	1.56
SSpro2 (SS)	73.71	1.00	74.57	2.44	69.64	2.39	74.55	1.26	67.43	1.58
PROFsec (PR)	74.42	0.90	70.72	2.77	70.70	2.63	73.87	1.33	69.80	1.50
Jpred	72.03	1.13	62.07	2.94	62.43	2.95	82.52	1.24	66.58	1.91
Consensus of PS, SA, SS	75.69	0.98	77.58	2.31	70.28	2.50	75.79	1.32	70.18	1.63
Consensus of PS, SA, PR	75.98	0.89	75.99	2.27	70.91	2.52	76.39	1.33	70.97	1.59
Consensus of PS, SS, PR	75.84	0.93	76.01	2.35	71.18	2.51	75.89	1.30	70.46	1.56
Consensus of SA, SS, PR	75.94	0.90	75.60	2.37	70.92	2.43	76.15	1.35	70.45	1.56
(b)										
PSIPRED	76.16	0.69	77.78	1.45	68.63	1.85	75.62	0.88	71.58	1.03
SSpro1	76.03	0.68	77.51	1.52	65.43	1.81	76.24	0.86	70.38	1.04
PROFsec	76.33	0.63	75.83	1.55	69.42	1.79	74.75	0.94	72.23	0.93
Jpred	74.63	0.63	68.24	1.76	60.82	1.89	82.51	0.83	69.08	1.04
Consensus	77.83	0.65	78.13	1.48	68.24	1.83	77.79	0.87	72.91	1.00

I metodi *consensus* incrementano l'affidabilità delle regioni predette in modo unanime.

Il Q_3 medio per queste regioni arriva al 82-84%, quello complessivo è attorno al 77-78%, superando i migliori metodi singoli di 1-2%. (*Albrecht & Tosatto, 2003*)

L'utilizzo dei metodi consenso migliora solo se si prendono i predittori migliori altrimenti il risultato peggiora !



STRUTTURA SECONDARIA IN PRATICA

1. I metodi singoli migliori (***PSIPRED, SSPRO***) arrivano a predire intorno al 77% il che significa che sbagliano nel 23% dei casi quindi ***nel singolo caso l'accuratezza può essere molto superiore o molto inferiore !***
2. Quanto detto nel punto 1 dipende da come sono stati allenati i metodi e dall'accuratezza del loro test set e da quanto la proteina che si sta studiando può essere un caso che rientra in quelli generali o quanto in realtà sia nuovo e precedentemente mai contemplato prima.
3. In genere si può ovviare a questo valutando il punteggio di affidabilità associato alla previsione per ogni singolo AA che i programmi danno come output, oppure utilizzare più metodi e confrontarli. Si possono trarre delle conclusioni in base al fatto che concordano o discordano. Può significare molte cose. Nel caso di discordanza può
 1. non essere affidabile la predizione;

o in certi casi:

 2. può essere sintomatico di una regione flessibile, che subisce degli switch conformazionali.

