

Corso di

Proprietà di Biopolimeri

Prof. R. Urbani

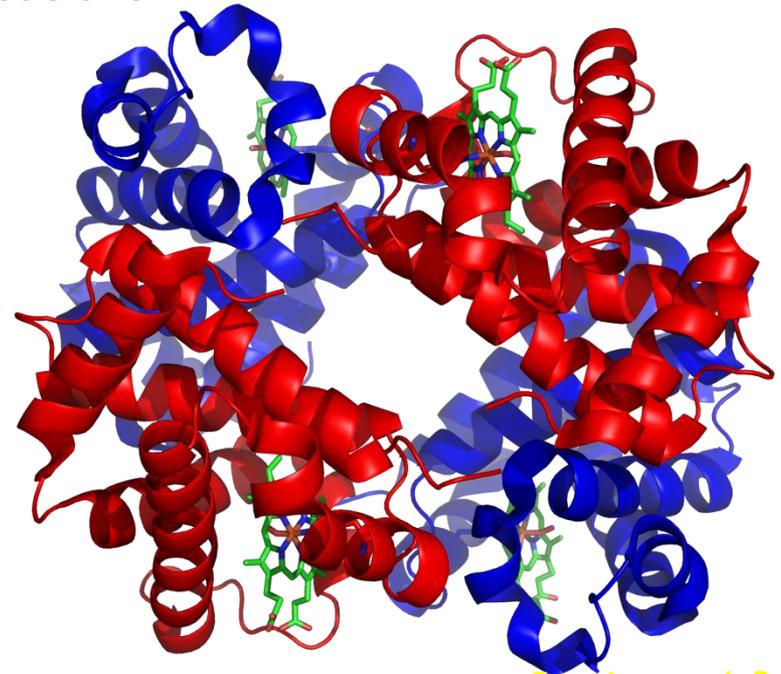
Dipartimento di Scienze Chimiche e Farmaceutiche

Tel 040 558 3684

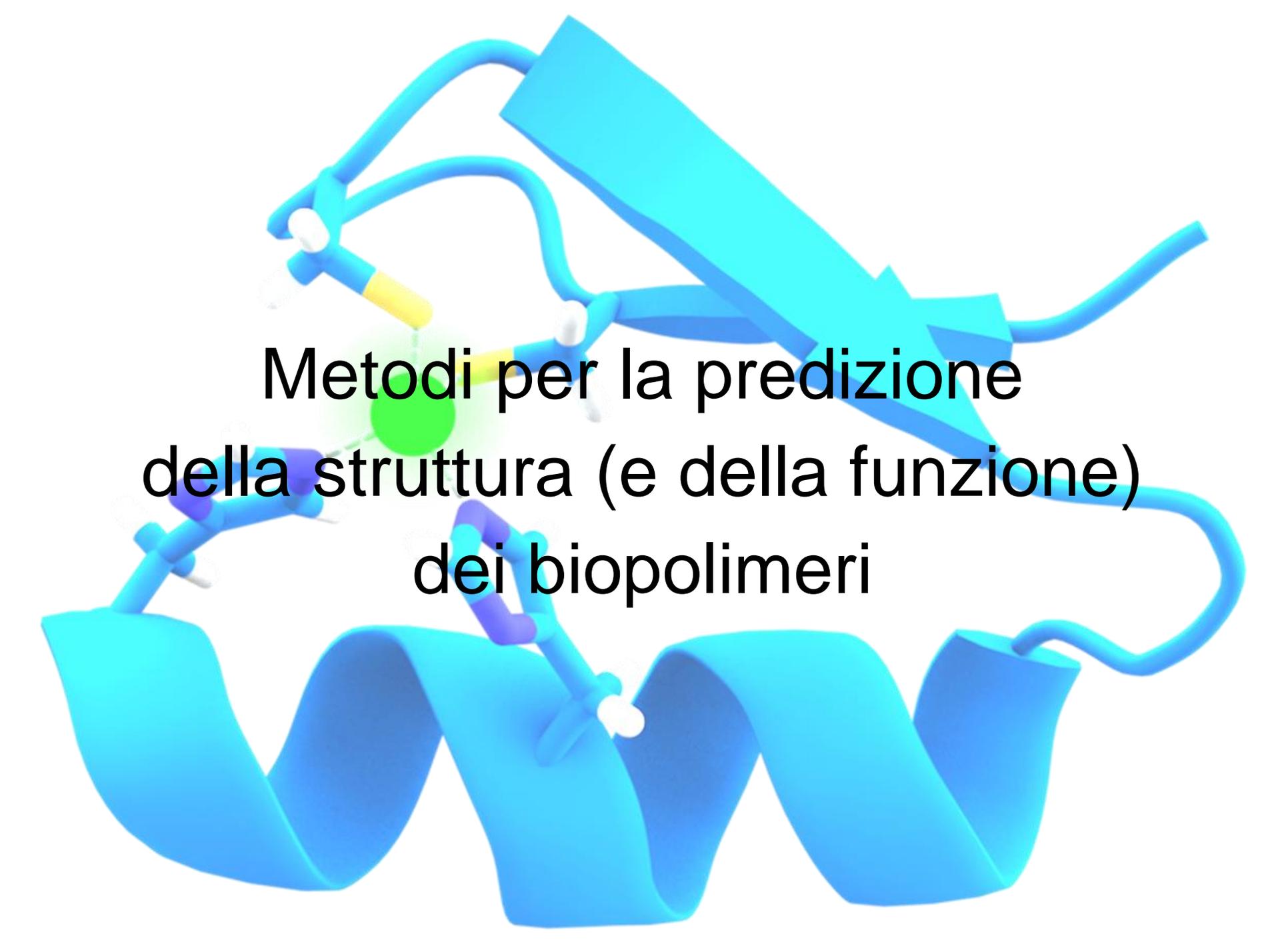
e-mail: rurbani@units.it

a.a. 2023-2024

Bioinformatica
Similitudine ed omologia



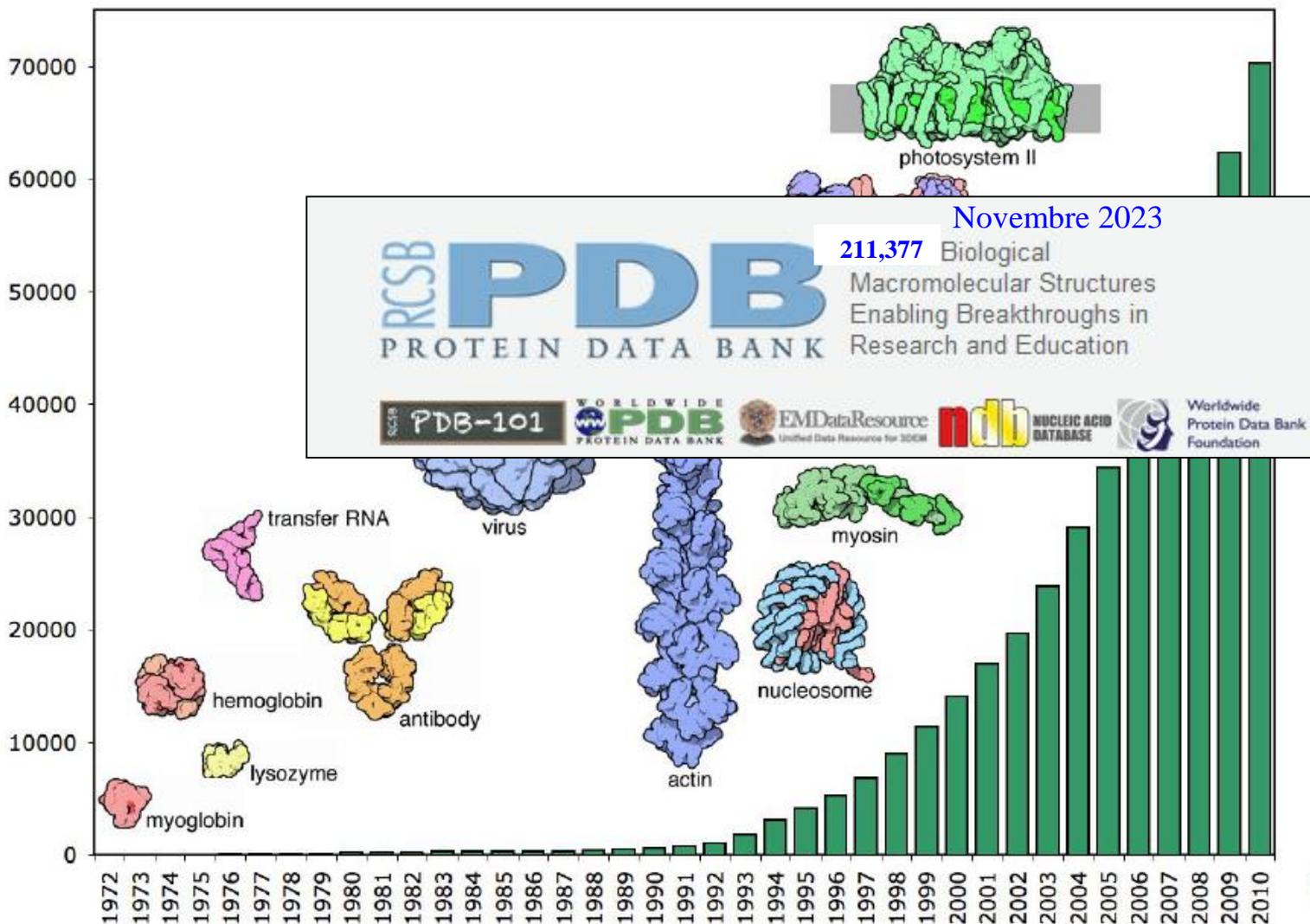
Lezione 4.2



Metodi per la predizione
della struttura (e della funzione)
dei biopolimeri

La struttura 3D di una proteina e' molto complessa (1958, John Kendrew, prima struttura della mioglobina)

strutture depositate nel Protein Data Bank



November 2021

183793 Biological
Macromolecular Structures
Enabling Breakthroughs in
Research and Education

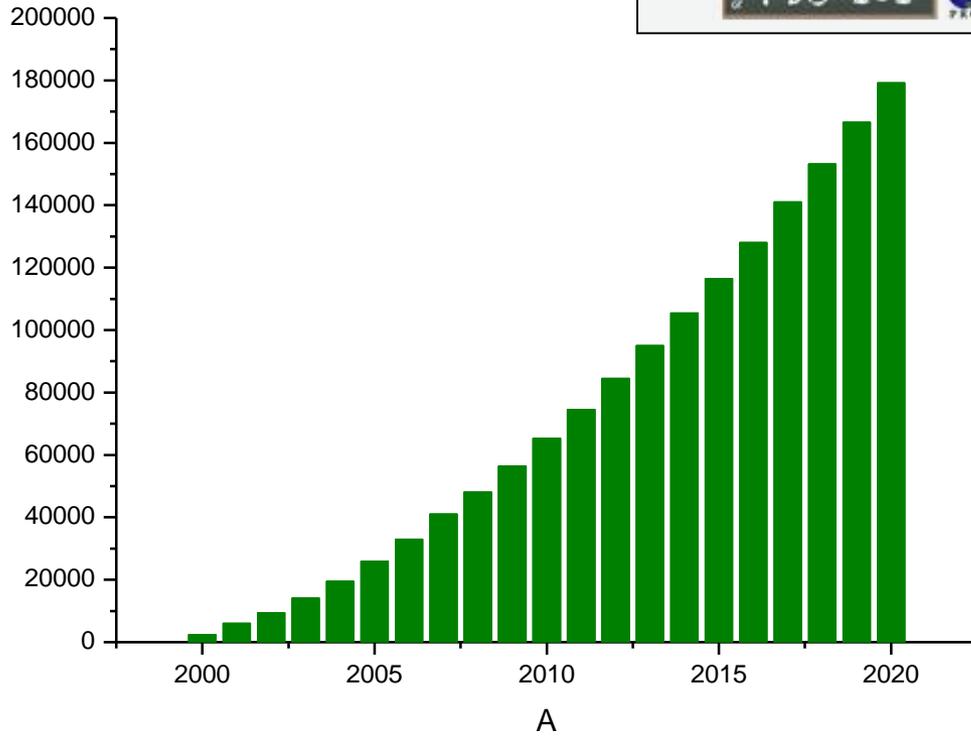


RCSB **PDB**
PROTEIN DATA BANK

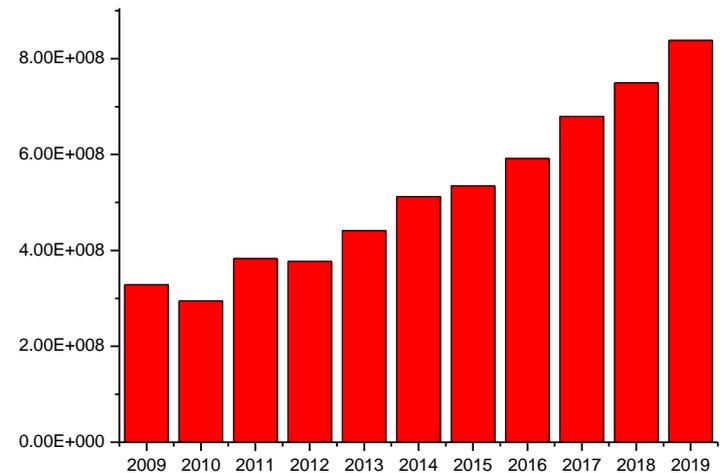








Download annuale



Lo scopo della **Bioinformatica** e della **Biochimica Computazionale** è quello di offrire strumenti e metodologie capaci di gestire ed analizzare la grande quantità di informazioni prodotte nel campo della ricerca biochimica,

determinata soprattutto **dall'enorme produzione** di sequenze di acidi nucleici e di proteine, che è il risultato degli studi relativi alle due discipline *omiche*: **genomica** e **proteomica**

gli strumenti computazionali e bioinformatici sono di grande aiuto, anche se essi devono essere considerati naturalmente complementari e non alternativi alle normali tecniche sperimentali

Dalla sequenza alla struttura secondaria

Dall'analisi delle sequenze delle proteine è possibile predire la struttura secondaria che tali sequenze possono assumere.

Metodi per la predizione delle strutture secondarie:

- **Approcci statistici**: Chou and Fasman, Garnier-Osguthorpe-Robson (GOR)
- **Proprietà chimico fisiche**: Rose, Eisenberg et al., ...
- **Riconoscimento di pattern**: Lim, Cohen et al., ...
- **Reti Neurali**: PHD, PSIPRED, ...

Calcolo dell'affidabilità delle predizioni:

Q3 score: la percentuale di residui di una proteina la cui struttura secondaria viene correttamente predetta dai vari metodi

Predizione strutture secondarie: accuratezza

La misura più usata per misurare il successo di un algoritmo di predizione di struttura secondaria è la misura di **accuratezza nei tre stati**: (eliche, beta, loop).

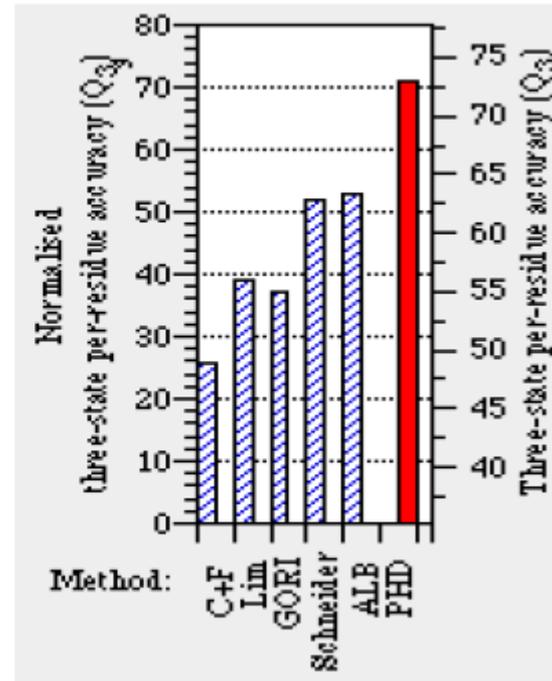
$$Q_3 = 100 \frac{\sum C_i}{N}$$

Sommatoria dei residui correttamente predetti in ciascun stato diviso i residui totali per 100.

Algoritmi di 1° generazione: <50% accuratezza

Algoritmi di 2° generazione: <60% accuratezza

Algoritmi di 3° generazione: >70% accuratezza



Algoritmi di 1° e 2° generazione: barre tratteggiate
Algoritmi di 3° generazione: barre rosse

METODI BASATI SULL'ANALISI DELLA SINGOLA SEQUENZA

Metodi di I° GENERAZIONE

- *Metodo Chou-Fasman (1974, $Q_3 \approx 50\%$)*

In realtà questi metodi calcolano la propensione alla struttura secondaria basandosi su singola sequenza e con un contesto locale molto ridotto

- *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 .

METODI DI III° GENERAZIONE:

Le RETI NEURALI

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.

$$Q_3 > 70\%$$

STRUMENTI DI PREVISIONE DELLE PROPRIETA' DI BIOPOLIMERI

METODI DI PRIMA GENERAZIONE

Scale di propensità

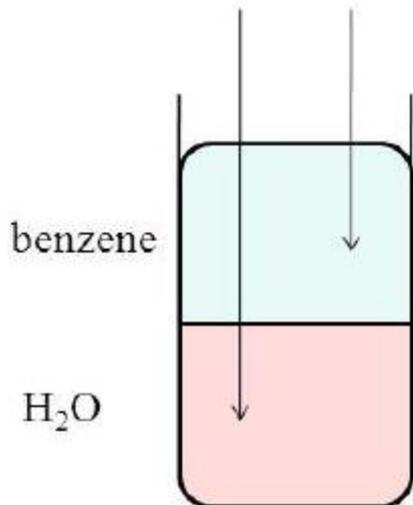
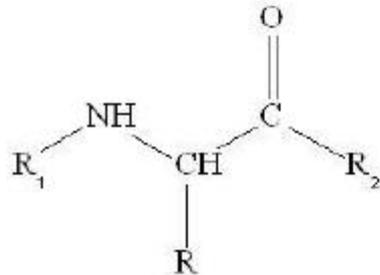
- Statistiche sulla presenza dei 20 amminoacidi nelle differenti strutture
- Considerazioni fisico-chimiche

Ad ogni tipo di amminoacido viene attribuito un valore di propensità ad assumere una certa struttura

An amino acid scale is defined by a numerical value assigned to each type of amino acid. The most frequently used scales are the hydrophobicity or hydrophilicity scales and the secondary structure conformational parameters scales, but many other scales exist which are based on different chemical and physical properties of the amino acids.

Molecular weight	Number of codon(s)
Bulkiness	Polarity / Zimmerman
Polarity / Grantham	Refractivity
Recognition factors	Hphob. / Eisenberg et al.
Hphob. OMH / Sweet et al.	Hphob. / Hopp & Woods
Hphob. / Kyte & Doolittle	Hphob. / Manavalan et al.
Hphob. / Abraham & Leo	Hphob. / Black
Hphob. / Bull & Breese	Hphob. / Fauchere et al.
Hphob. / Guy	Hphob. / Janin
Hphob. / Miyazawa et al.	Hphob. / Rao & Argos
Hphob. / Roseman	Hphob. / Tanford
Hphob. / Wolfenden et al.	Hphob. / Welling & al
Hphob. HPLC / Wilson & al	Hphob. HPLC / Parker & al
Hphob. HPLC pH3.4 / Cowan	Hphob. HPLC pH7.5 / Cowan
Hphob. / Rf mobility	HPLC / HFBA retention
HPLC / TFA retention	Transmembrane tendency
HPLC / retention pH 2.1	HPLC / retention pH 7.4
% buried residues	% accessible residues
Hphob. / Chothia	Hphob. / Rose & al
Ratio hetero end/side	Average area buried
Average flexibility	alpha-helix / Chou & Fasman
beta-sheet / Chou & Fasman	beta-turn / Chou & Fasman
alpha-helix / Deleage & Roux	beta-sheet / Deleage & Roux
beta-turn / Deleage & Roux	Coil / Deleage & Roux
alpha-helix / Levitt	beta-sheet / Levitt
beta-turn / Levitt	Total beta-strand
Antiparallel beta-strand	Parallel beta-strand
A.A. composition	A.A. comp. in Swiss-Prot
Relative mutability	

Le scale di idropatia



costante di ripartizione

$$K_R = \frac{[\text{amminoacido}]_{\text{benzene}}}{[\text{amminoacido}]_{\text{H}_2\text{O}}}$$

L'energia libera per il trasferimento dell'amminoacido dall'acqua al benzene è una misura della idrofobicità

$$\Delta G_{\text{H}_2\text{O} \rightarrow \text{benzene}} = RT \ln K_R$$

scala
Kyte & Doolittle (1982)

Ala	1.8
Arg	-4.5
Asn	-3.5
Asp	-3.5
Cys	2.5
Gln	-3.5
Glu	-3.5
Gly	-0.4
His	-3.2
Ile	4.5
Leu	3.8
Lys	-3.9
Met	1.9
Phe	2.8
Pro	-1.6
Ser	-0.8
Thr	-0.7
Trp	-0.9
Tyr	-1.3
Val	4.2



Swiss Institute of
Bioinformatics

www.expasy.org

Home

About

SIB News

Contact

Expasy

Swiss Bioinformatics Resource Portal



e.g. [BLAST](#), [UniProt](#), [MSH6](#), [Albumin...](#)

 **Genes & Genomes**

- Genomics
- Metagenomics
- Transcriptomics

 **Proteins & Proteomes**

 **Evolution & Phylogeny**

- Evolution biology
- Population genetics

 **Structural Biology**

- Drug design
- Medicinal chemistry
- Structural analysis

 **Systems Biology**

- Glycomics
- Lipidomics
- Metabolomics

 **Text mining & Machine learning**

SIB Resources ⓘ

Other Resources of SIB Groups



SwissParam

Topology and parameters for small molecules



SWISS-MODEL Workspace

Fully automated protein structure homology-modeling server



SwissBiolsostere

Database of molecular replacements for ligand design



OpenStructure

Molecular modelling and visualisation environment



Swiss-PdbViewer

Display, analyse and superimpose protein 3D structures



MARCOIL

Coiled-coil regions in protein sequences



SwissSidechain

Database of non-natural amino acid side chains



Click2Drug

Directory of computational drug design tools



QMEAN

Protein model quality estimation



rBAN

Mapping of non-ribosomal peptides in SMILES format



SwissTargetPrediction

Target prediction for bioactive small molecules

1.2 Analisi del profilo di idrofobicità

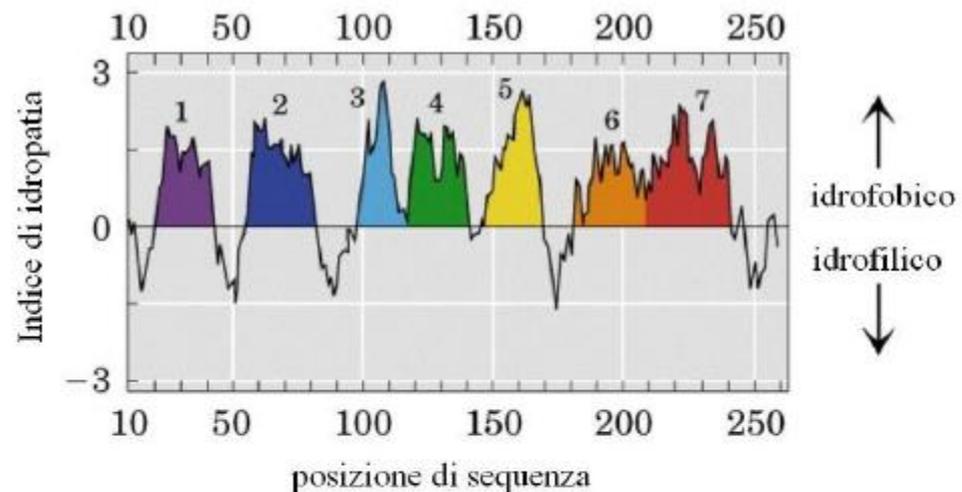
La conoscenza della struttura primaria di una proteina permette anche di prevederne il profilo di idrofobicità, vale a dire come cambiano le caratteristiche di idrofilicità e di idrofobicità lungo la sequenza. Questa informazione può essere utilizzata per prevedere la possibile interazione di una proteina con le membrane fosfolipidiche della cellula.

Amino Acid Hydropathy Scores [1]

Amino Acid	One Letter Code	Hydropathy Score
Isoleucine	I	4.5
Valine	V	4.2
Leucine	L	3.8
Phenylalanine	F	2.8
Cysteine	C	2.5
Methionine	M	1.9
Alanine	A	1.8
Glycine	G	-0.4
Threonine	T	-0.7
Serine	S	-0.8
Tryptophan	W	-0.9
Tyrosine	Y	-1.3
Proline	P	-1.6
Histidine	H	-3.2
Glutamic acid	E	-3.5
Glutamine	Q	-3.5
Aspartic acid	D	-3.5
Asparagine	N	-3.5
Lysine	K	-3.9
Arginine	R	-4.5

Kyte J, Doolittle RF (May 1982).

"A simple method for displaying the hydropathic character of a protein".
J. Mol. Biol. 157 (1): 105-32.

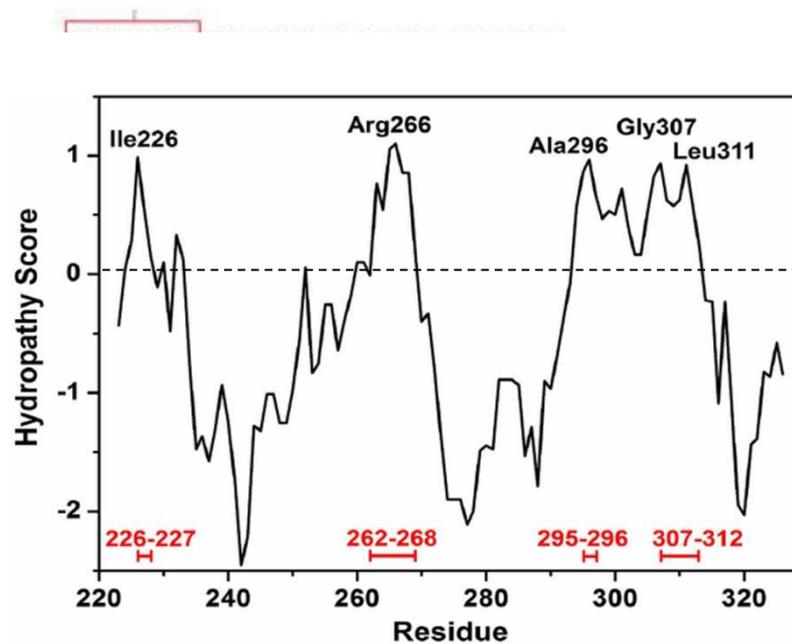


Nella parte alta della schermata che otterrete vengono indicati i valori di idrofobicità/idrofilicità attribuiti a ciascun residuo, valori più positivi indicano maggiore idrofobicità. I picchi del grafico sottostante rappresentano quindi le regioni di maggiore idrofobicità (vale a dire quelle in cui più residui idrofobici si susseguono in sequenza) e potrebbero corrispondere a regioni transmembrana.

Plot di idropatia

- Valori di idrofobicità di Kyte & Dolittle.

Ala: 1.800
 Arg: -4.500
 Asn: -3.500
 Asp: -3.500
 Cys: 2.500
 Gln: -3.500
 Glu: -3.500
 Gly: -0.400
 His: -3.200
 Ile: 4.500
 Leu: 3.800
 Lys: -3.900
 Met: 1.900
 Phe: 2.800
 Pro: -1.600
 Ser: -0.800
 Thr: -0.700
 Trp: -0.900
 Tyr: -1.300
 Val: 4.200

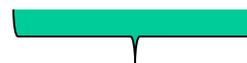


Esempio: HELP (Human Elastic Like Polymers)

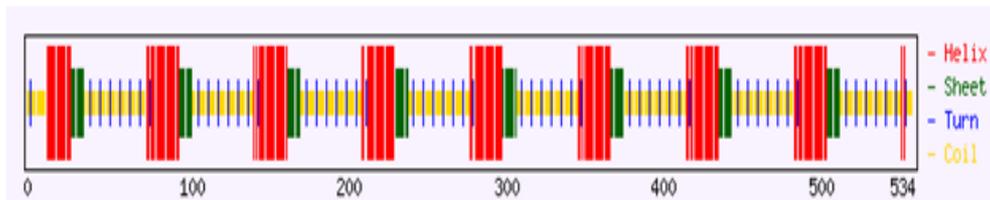
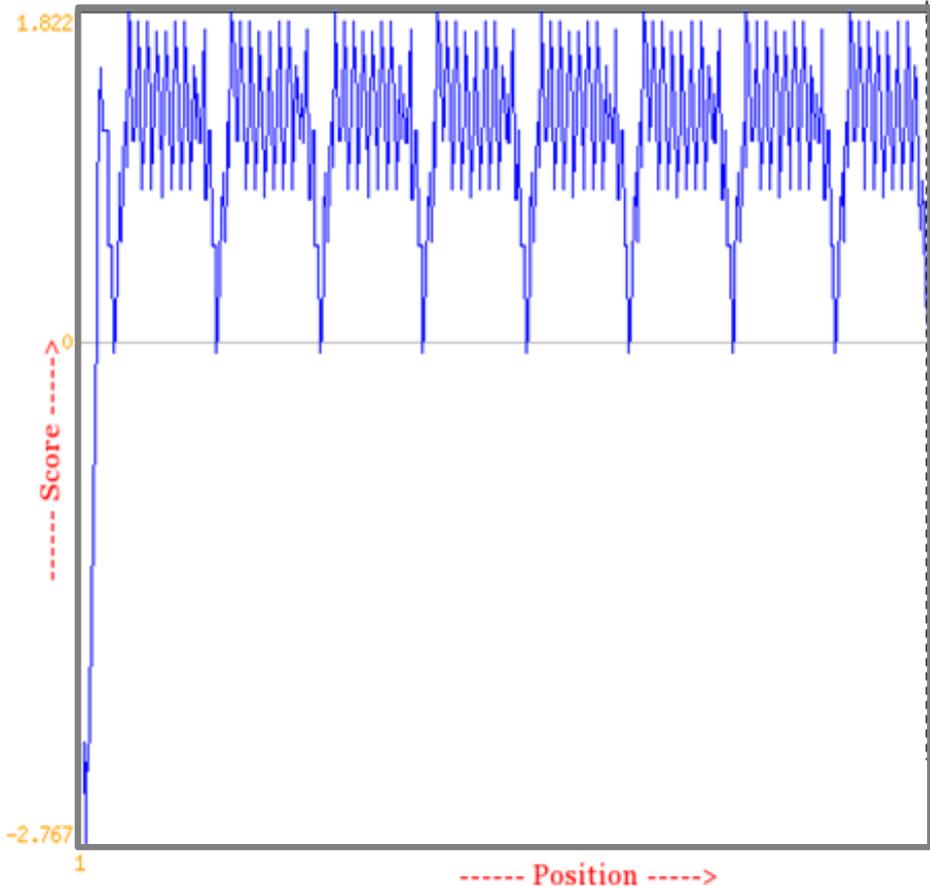
Target Sequence:

MRGSHHHHHH¹⁰ GSAAAAAAA²⁰ KAAAKAAQFG³⁰ LVPGVGVAPG⁴⁰ VGVAPGVGV⁵⁰ PGVGLAPGVG⁶⁰ VAPGVGVAPG⁷⁰
VG⁸⁰VAPGIAP⁹⁰AAAAAKAAAK¹⁰⁰AAQFGLVPGV¹¹⁰GVAPGVGVAP¹²⁰GVGVAPGVGL¹³⁰APGVGVAPGV¹⁴⁰GVAPGVGVAP¹⁵⁰
GIAPAAAAAA¹⁶⁰KAAAKAAQFG¹⁷⁰LVPGVGVAPG¹⁸⁰VG¹⁹⁰VAPGVGV²⁰⁰VAPGVGVAPG²¹⁰VG²²⁰VAPGIAP²³⁰AAAAAKAAAK²⁴⁰AAQFGLVPGV²⁵⁰GVAPGVGVAP²⁶⁰GVGVAPGVGL²⁷⁰APGVGVAPGV²⁸⁰GIAPAAAAAA²⁹⁰
KAAAKAAQFG³⁰⁰LVPGVGVAPG³¹⁰VG³²⁰VAPGVGV³³⁰VAPGVGVAPG³⁴⁰VG³⁵⁰VAPGIAP³⁶⁰AAAAAKAAAK³⁷⁰AAQFGLVPGV³⁸⁰GVAPGVGVAP³⁹⁰GVGVAPGVGL⁴⁰⁰APGVGVAPGV⁴¹⁰GVAPGVGVAP⁴²⁰GIAPAAAAAA⁴³⁰KAAAKAAQFG⁴⁴⁰
LVPGVGVAPG⁴⁵⁰VG⁴⁶⁰VAPGVGV⁴⁷⁰VAPGVGVAPG⁴⁸⁰VG⁴⁹⁰VAPGIAP⁵⁰⁰AAAAAKAAAK⁵¹⁰AAQFGLVPGV⁵²⁰GVAPGVGVAP⁵³⁰GVGVAPGVGL⁵⁴⁰APGVGVAPGV⁵⁵⁰GVAPGVGVAP⁵⁶⁰GIAP

-VAPGVG-



Kyte & Doolittle Hydrophobicity Plot



Number of amino acids: 534
Molecular weight: 44729.60
Theoretical pI: 11.68

Amino acid composition:

Ala (A)	154	28.8%
Arg (R)	1	0.2%
Asn (N)	0	0.0%
Asp (D)	0	0.0%
Cys (C)	0	0.0%
Gln (Q)	8	1.5%
Glu (E)	0	0.0%
Gly (G)	130	24.3%
His (H)	6	1.1%
Ile (I)	8	1.5%
Leu (L)	16	3.0%
Lys (K)	16	3.0%
Met (M)	1	0.2%
Phe (F)	8	1.5%
Pro (P)	72	13.5%
Ser (S)	2	0.4%
Thr (T)	0	0.0%
Trp (W)	0	0.0%
Tyr (Y)	0	0.0%
Val (V)	112	21.0%
Pyl (O)	0	0.0%
Sec (U)	0	0.0%

Total number of negatively charged residues (Asp + Glu): 0
Total number of positively charged residues (Arg + Lys): 17

Total number of atoms: 6505

Extinction coefficients:

As there are no **Trp**, **Tyr** or **Cys** in the region considered, your protein should not be visible by UV spectrophotometry.

Grand average of hydropathicity (GRAVY): 1.097

Table 1. Chemico-physical parameters obtained using ExPASy Tools (ProtParam on-line software)



	MW	<u>p.I.</u>	<u>Hydropathy Index (GRAVY)</u>	% polar <u>a.a.</u>	% charged <u>a.a.</u>	% aromaticity
HUG	60,406	9.9	0.77	5.3	10.0	3.1
HELP	44,886	11.7	1.1	1.9	3.2	0
Human elastin	66,135	10.4	0.62	3.3	7.3	5.5

Metodo di Chou e Fassman

Una delle tecniche di predizione della struttura secondaria su base statistica più usate è quella elaborata da **Chou e Fassman (1974)**, che va a valutare la propensione di ciascun amminoacido a trovarsi in una particolare struttura secondaria (elica, β -strand e coil)



Questo metodo fornisce una tabella nella quale ciascun amminoacido viene classificato con un coefficiente, che riflette la frequenza con la quale esso **forma, interrompe o è indifferente** alla formazione di ciascun tipo di struttura secondaria (“former”, “breaker” o “indifferent”).

.



Si assegna quindi ad ogni residuo la conformazione avente **maggiore probabilità media** su una finestra di un certo numero di amminoacidi (da 5 a 7) che lo circondano.

Previsione della struttura secondaria secondo Chou e Fassman

Il principio della previsione e' il seguente:

- Analizzare le strutture tridimensionali di proteine di riferimento (erano 15 nell'articolo originale) a partire da dati di diffrazione dei raggi X.
- Identificare gli amminoacidi che fanno parte delle tre strutture secondarie principali: α -elica, foglietto β , ripiegamento β .

- Calcolare le frequenze con le quali ciascun residuo ricorre in un'elica

$f_{\alpha} = n_{\alpha} / n_T$ con n_{α} : numero di volte che il residuo ricorre in un'elica;
 n_T numero totale delle volte che il residuo compare nella
popolazione di proteine scelta),

$f_{\beta} = n_{\beta} / n_T$ in un foglietto

$f_{\tau} = n_{\tau} / n_T$ in un ripiegamento

•Calcolare il valor medio di f_{α} , f_{β} e f_{τ} per i 20 residui, con la formula seguente: $\langle f_{\alpha} \rangle = \Sigma f_{\alpha} / 20$.

•Calcolare la propensione a formare l'una o l'altra delle strutture secondarie:

$$P_{\alpha} = f_{\alpha} / \langle f_{\alpha} \rangle; \quad P_{\beta} = f_{\beta} / \langle f_{\beta} \rangle; \quad P_{\tau} = f_{\tau} / \langle f_{\tau} \rangle.$$

•I valori di P indicano la propensione del residuo ad essere in una delle strutture secondarie quando il valore è superiore alla media (cioè $P > 1$).

Conformational Preferences of the Amino Acids

Amino acid	Preference		
	α -helix	β -strand	Reverse turn
Glu	1.59	0.52	1.01
Ala	1.41	0.72	0.82
Leu	1.34	1.22	0.57
Met	1.30	1.14	0.52
Gln	1.27	0.98	0.84
Lys	1.23	0.69	1.07
Arg	1.21	0.84	0.90
His	1.05	0.80	0.81
Val	0.90	1.87	0.41
Ile	1.09	1.67	0.47
Tyr	0.74	1.45	0.76
Cys	0.66	1.40	0.54
Trp	1.02	1.35	0.65
Phe	1.16	1.33	0.59
Thr	0.76	1.17	0.90
Gly	0.43	0.58	1.77
Asn	0.76	0.48	1.34
Pro	0.34	0.31	1.32
Ser	0.57	0.96	1.22
Asp	0.99	0.39	1.24

I valori di P per in singoli amminoacidi e per le strutture secondarie sono riportati nella tabella seguente:

	P_{α}			P_{β}			P_{τ}
Glu	1,51	(HA)	Val	1,70	(HB)	Asn	1,56
Met	1,45	(HA)	Ile	1,60	(HB)	Gly	1,56
Ala	1,42	(HA)	Tyr	1,47	(HB)	Pro	1,52
Leu	1,21	(HA)	Phe	1,38	(hB)	Asp	1,46
Lys	1,16	(hA)	Trp	1,37	(hB)	Ser	1,43
Phe	1,13	(hA)	Leu	1,30	(hB)	Cys	1,19
Gln	1,11	(hA)	Cys	1,19	(hB)	Tyr	1,14
Trp	1,08	(hA)	Thr	1,19	(hB)	Lys	1,01
Ile	1,08	(hA)	Gln	1,10	(hB)	Gln	0,98
Val	1,06	(hA)	Met	1,05	(hB)	Thr	0,98
Asp	1,01	(IA)	Arg	0,93	(iB)	Trp	0,96
His	1,00	(IA)	Asn	0,89	(iB)	Arg	0,96
Arg	0,98	(iA)	His	0,87	(iB)	His	0,95
Thr	0,83	(iA)	Ala	0,83	(iB)	Glu	0,74
Ser	0,77	(iA)	Ser	0,75	(bB)	Ala	0,66
Cys	0,70	(iA)	Gly	0,75	(bB)	Met	0,60
Tyr	0,69	(bA)	Lys	0,74	(bB)	Phe	0,60
Asn	0,67	(bA)	Pro	0,55	(BB)	Leu	0,59
Pro	0,57	(BA)	Asp	0,54	(BB)	Val	0,50
Gly	0,57	(BA)	Glu	0,37	(BB)	Ile	0,47

HA, hA, IA, iA: induttori di α -eliche (forte, medio, debole, indifferente)

bA, BA: interruttori di α -eliche (deboli e forti)

HB, hB, IB, iB: induttori di foglietto β (forte, medio, debole, indifferente)

bB, BB: interruttori di foglietto β (debole e forte)

Il dataset originale comprendeva solo 15 proteine; in seguito venne ampliato fino a 144 proteine.

L'affidabilità del metodo è abbastanza bassa (circa **50%**), tuttavia il metodo Chou-Fasman è ancora molto utilizzato grazie soprattutto alla semplicità di approccio.

Struttura secondaria: Metodo di Chou-Fasman

Dato un insieme di strutture note, si conta quante volte ognuno degli amminoacidi è presente in una data struttura e si determina il grado di indipendenza tra l'amminoacido e la struttura

Esempio:

ALAKSLAKPSDTLAKSDFREKWEWLKLLKALACCKLSAAL
hhhhhhhhccccccccccccchhhhhhhhhhhhhhhhh

$$\begin{aligned} N(A,h) &= 7, & N(A,c) &= 1, & N &= 40, & N(A) &= 8, & N(h) &= 27 \\ P(A,h) &= 7/40, & P(A) &= 8/40, & P(h) &= 27/40 \end{aligned}$$

Se amminoacido e struttura sono indipendenti:

$$P(A,h) = P(A)P(h)$$

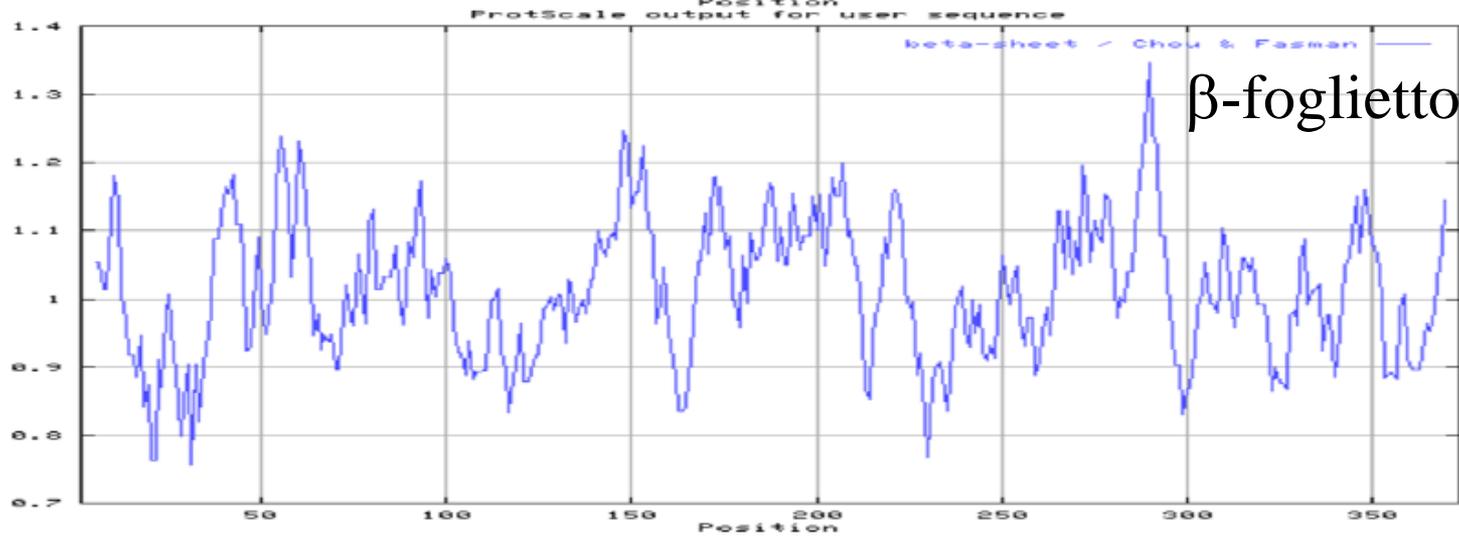
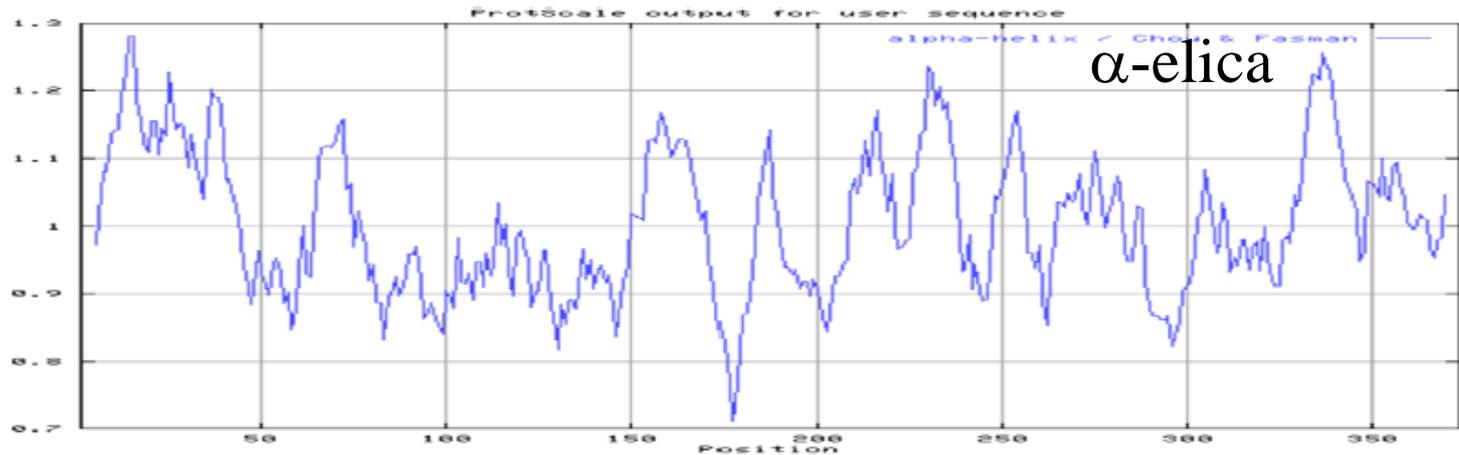
Il rapporto $P(A,h)/P(A)P(h)$ è detto propensione

1.29

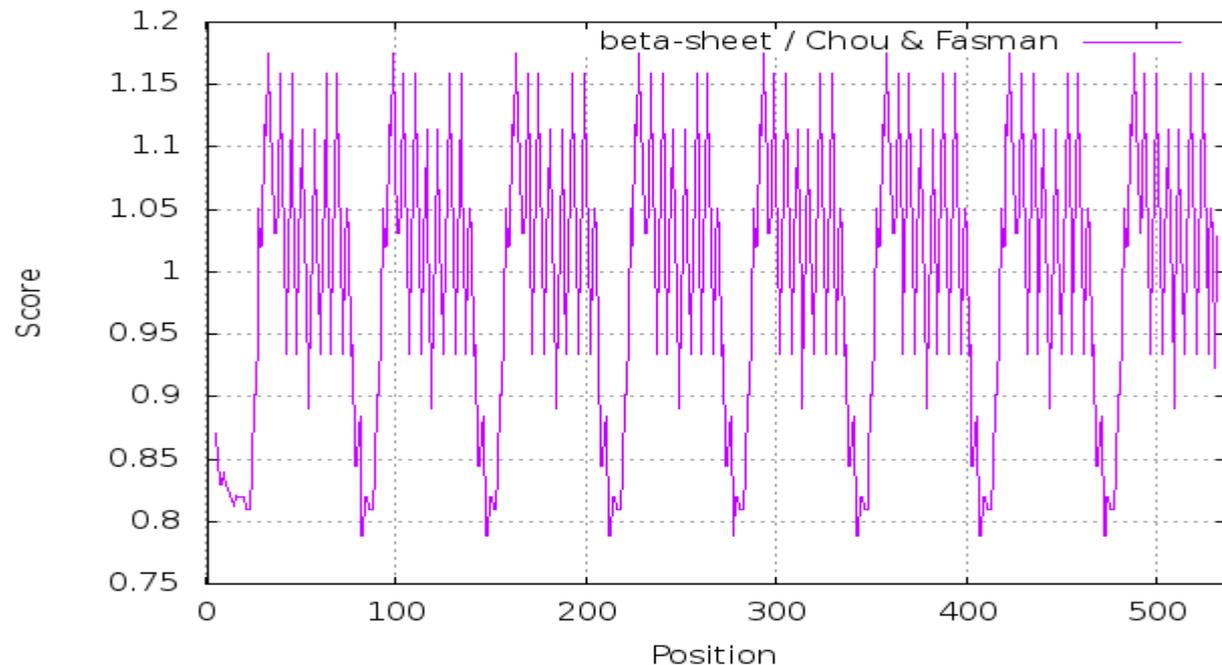
RIFERIMENTI BIBLIOGRAFICI

Chou,P,Y., Fasman, G,D, 1973; *J, Mol, Biol.*, 74, 263-81,
Chou,P,Y., Fasman, G,D, 1974; *Biochemistry*, 13, 211-22,
Chou,P,Y., Fasman, G,D, 1974; *Biochemistry*, 13, 222-45

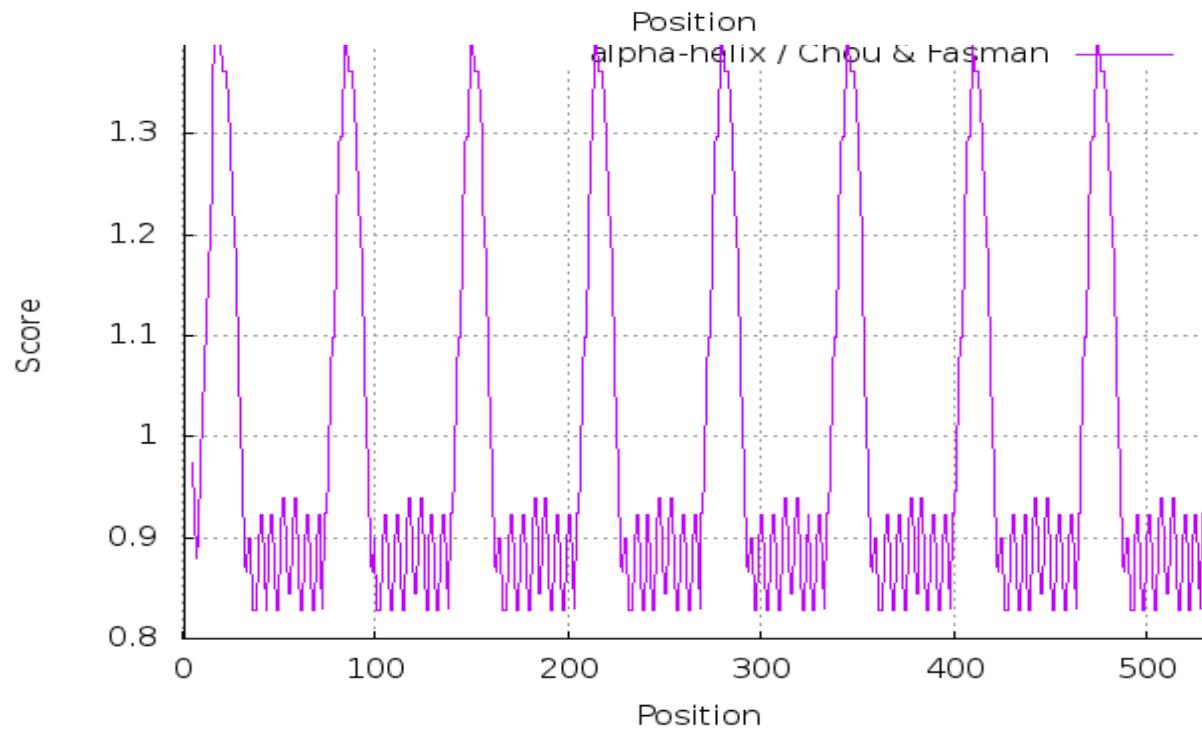
Data una nuova sequenza si mettono in grafico i valori di propensità, residuo per residuo e si ricava una predizione di struttura secondaria



ProtScale output for user_sequence

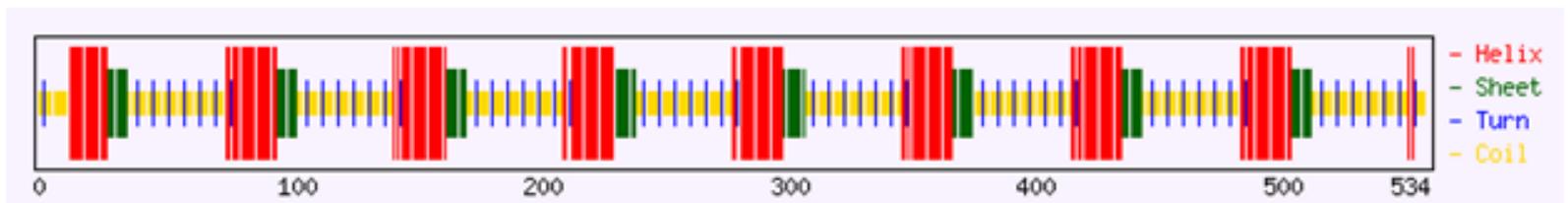


HELP



Struttura secondaria di HELP

Query 1	MRGSHHHHHHGSAAAAAAKAAAKAAQFGLVPGVGVAPGVGVAPGVGVAPGVGLAPGVGVAPGVGVAPG	70
Helix 1	HHHHHHHHHHHHHHHHHHHH	70
Sheet 1	EEEEEEEE	70
Turns 1	T T T T T T T	70
Struc 1	CCTCCCCCCCCHHHHHHHHHHHHHHEEEEEEECCCTCCCCCTCCCCCTCCCCCTCCCCCTCCCCCTC	70
Query 71	VGVPAGIAPAAAAAAKAAAKAAQFGLVPGVGVAPGVGVAPGVGVAPGVGLAPGVGVAPGVGVAPGVGVAP	140
Helix 71	HHHHHHHHHHHHHHHHHHHHHH HHH	140
Sheet 71	EEEEEEEE	140
Turns 71	T T T T T T T	140
Struc 71	CCHHTHHHHHHHHHHHHHHHHHEEEEEEECCCTCCCCCTCCCCCTCCCCCTCCCCCTCCCCCTCCCHHT	140
Query 141	GIAPAAAAAAKAAAKAAQFGLVPGVGVAPGVGVAPGVGVAPGVGLAPGVGVAPGVGVAPGVGVAPGIAPA	210
Helix 141	HHHHHHHHHHHHHHHHHHHH HHHHHHHH	210
Sheet 141	EEEEEEEE	210
Turns 141	T T T T T T T	210
Struc 141	HHHHHHHHHHHHHHHHHEEEEEEECCCTCCCCCTCCCCCTCCCCCTCCCCCTCCCCCTCCCHHTHHHHHH	210
Query 211	AAAAAKAAAKAAQFGLVPGVGVAPGVGVAPGVGVAPGVGLAPGVGVAPGVGVAPGVGVAPGIAPAAAAAA	280
Helix 211	HHHHHHHHHHHHHHHHHH HHHHHHHHHHHHH	280
Sheet 211	EEEEEEEE	280
Turns 211	T T T T T T T	280
Struc 211	HHHHHHHHHHHHHEEEEEEECCCTCCCCCTCCCCCTCCCCCTCCCCCTCCCCCTCCCHHTHHHHHHHHHH	280
Query 281	KAAAKAAQFGLVPGVGVAPGVGVAPGVGVAPGVGLAPGVGVAPGVGVAPGVGVAPGIAPAAAAAAKAAAK	350
Helix 281	HHHHHHHHHHHH HHHHHHHHHHHHHHHHHHH	350
Sheet 281	EEEEEEEE	350
Turns 281	T T T T T T T	350
Struc 281	HHHHHHHEEEEEEECCCTCCCCCTCCCCCTCCCCCTCCCCCTCCCCCTCCCHHTHHHHHHHHHHHHHH	350
Query 351	AAQFGLVPGVGVAPGVGVAPGVGVAPGVGLAPGVGVAPGVGVAPGVGVAPGIAPAAAAAAKAAAKAAQFG	420
Helix 351	HHHHH HHHHHHHHHHHHHHHHHHHHHHHHH	420
Sheet 351	EEEEEEEE EEE	420
Turns 351	T T T T T T T	420
Struc 351	HHEEEEEEECCCTCCCCCTCCCCCTCCCCCTCCCCCTCCCCCTCCCHHTHHHHHHHHHHHHHHHEEE	420



	number of a.a.	α -helix %	β -strand %	random coil + β -turn, %
HELP	536	26	4	70
Human ELP	757	23	12	65

A cosa può servire il risultato della predizione della struttura secondaria ?

L'utilizzo dipende dall'affidabilità della predizione:

- definizione della classe strutturale e confronto con classificazione di proteine (db SCOP, CATH)
- confronto con organizzazione di struttura secondaria di proteine note
- confronto con risultati di altri metodi (anche metodi di predizione della struttura terziaria)



[About](#)

[Contact](#)

[Download](#)

Il database SCOP mira a fornire una descrizione dettagliata e completa delle relazioni strutturali ed evolutive tra le proteine la cui struttura tridimensionale è nota e depositata nella Protein Data Bank.

CATH / Gene3D v4.3

151 million protein domains classified into 5,841 superfamilies



Il database CATH è una risorsa online gratuita e disponibile al pubblico che fornisce informazioni sulle relazioni evolutive dei domini proteici.

È stato creato a metà degli anni '90 presso l'University College di Londra.

Metodi di predizione della struttura secondaria delle proteine:

Metodi di Chou-Fasman si basa sull'analisi statistica della composizione in residui delle strutture secondarie presenti nella PDB.

(http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=misc1)

GOR si basa sull'analisi statistica della composizione in residui delle strutture secondarie presenti nella PDB.

(http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html)

AGADIR per predire la percentuale di residui in elica

(<http://www.embl-heidelberg.de/Services/serrano/agadir/agadir-start.html>)

PHD prende in input o una sequenza o un allineamento multiplo ed usa le reti neurali.

(<http://www.embl-heidelberg.de/predictprotein/predictprotein.html>)

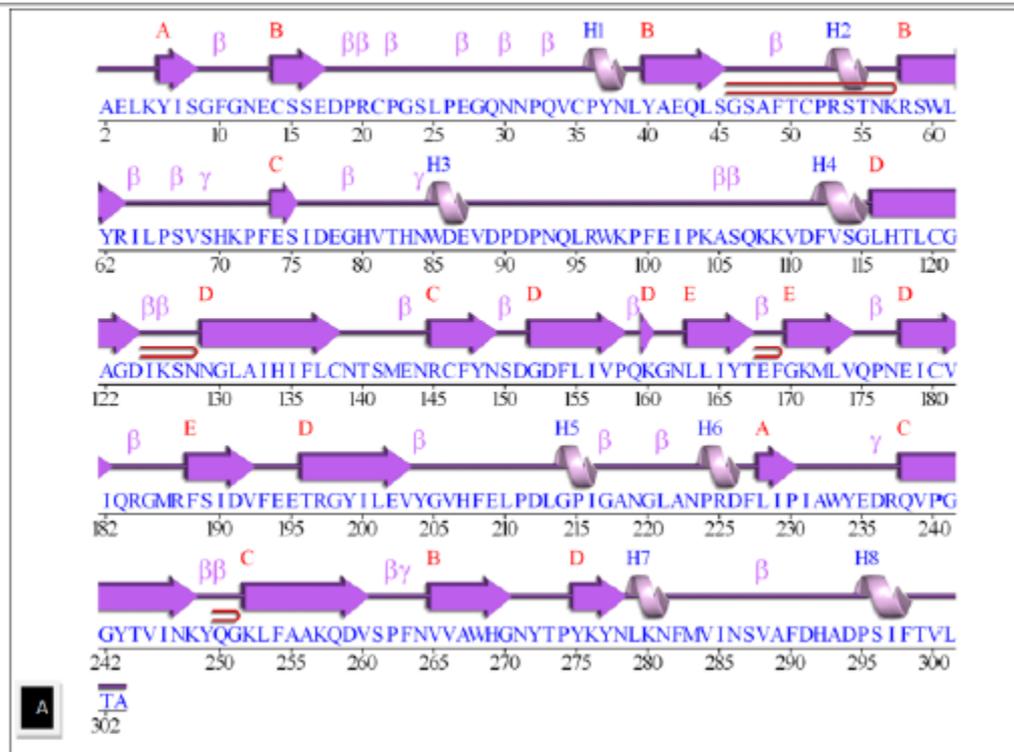
PSIPRED utilizza un sistema di due reti neurali. (<http://bioinf.cs.ucl.ac.uk/psipred/>)

PREDATOR si basa sull'applicazione del metodo del k-esimo vicino che usa le reti neurali

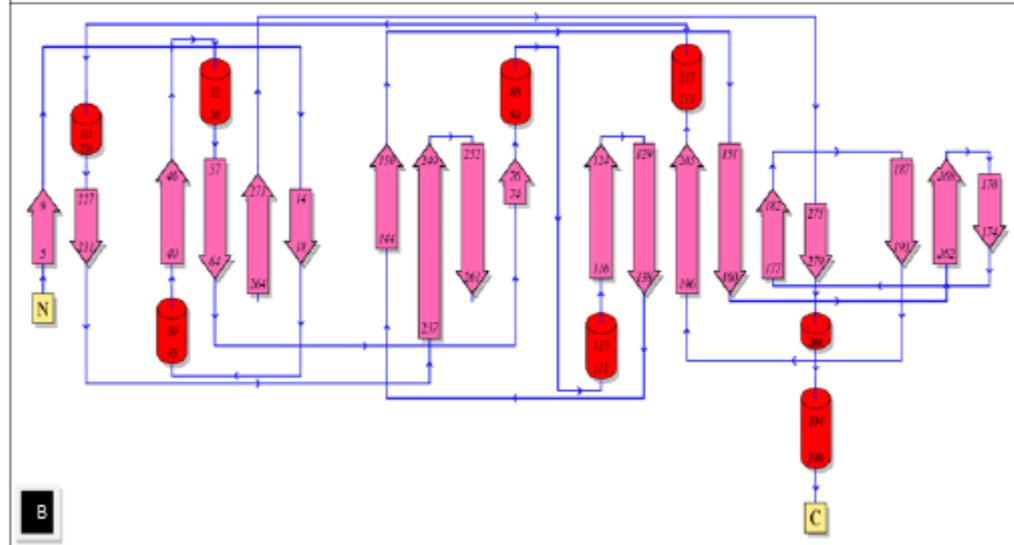
(<http://bioweb.pasteur.fr/seqanal/interfaces/predator-simple.html>)

JPRED (<http://www.compbio.dundee.ac.uk/Software/JPred/jpred.html>) fa un consensus di vari metodi

Struttura secondaria



Struttura terziaria



Il metodo GOR (Garnier-Osguthorpe-Robson, 1978)

GOR si basa sull'analisi statistica della composizione in residui delle strutture secondarie presenti in PDB.

Utilizza una finestra di **17 residui (8-1-8)** per determinare la probabilità del residuo centrale di far parte di una specifica struttura secondaria



Utilizzando un set di proteine a struttura nota, vengono calcolate le frequenze con le quali un certo aminoacido, in presenza di altri aminoacidi vicini, si trovi ad assumere una certa conformazione (alpha, beta o loops) e fornisce una matrice di punteggio per ciascuna struttura.

SOLUZIONE PROPOSTA
METODI DI III° GENERAZIONE:
Le RETI NEURALI

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.

$$Q_3 > 70\%$$

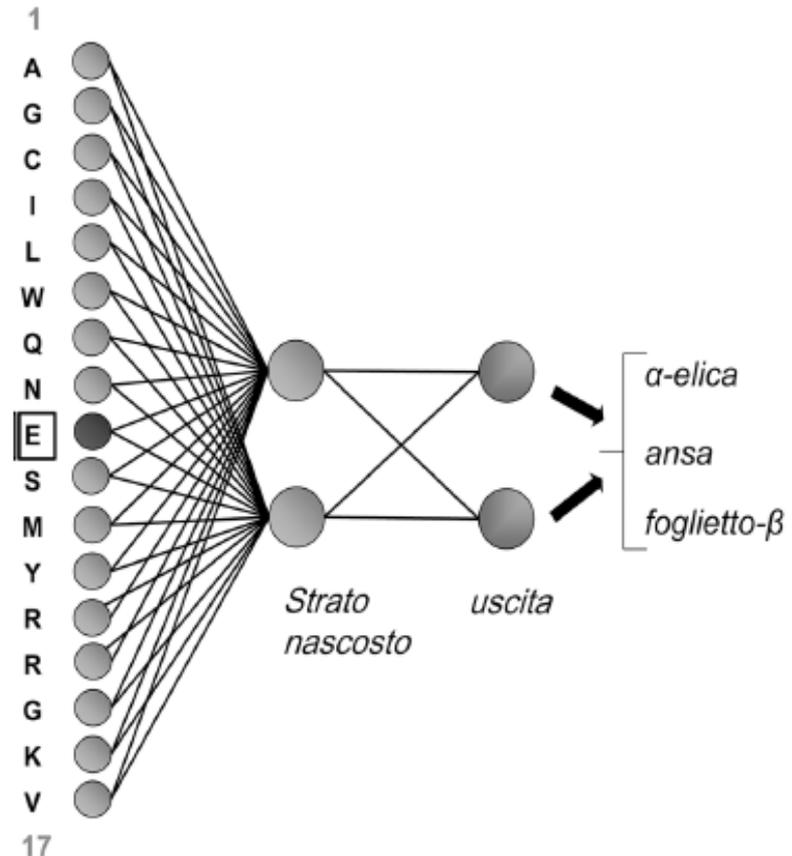
Predittori di 3° generazione: reti neurali

Sono su **reti neurali**, sistemi che simulano le operazioni dei neuroni del cervello.

Di solito vi sono 3 layer, ciascuno composto da unità elementari (neuroni) che formano connessioni (sinapsi) con ciascun neurone del layer successivo

I segnali passano da un **input layer** che raccoglie informazioni su una finestra di residui, sono inviati ad un **hidden layer** che elabora i segnali secondo una funzione matematica e li invia ad un **output layer** che predice la struttura secondaria dell'ammino acido centrale nella finestra.

I segnali sono pesati e tali pesi possono essere modificati durante l'operazione di **apprendimento** in cui vengono minimizzati gli errori di predizione in un **training set** di proteine a struttura nota



Cosa c'entra tutto ciò con le proteine?

le reti neurali possono **riconoscere** la presenza di *particolari segnali di sequenza* e prevedere proprietà strutturali e funzionali associate al segnale. Esempio: la **struttura secondaria**

L'idea:

la rete può “leggere” una porzione di sequenza e decidere se il residuo centrale appartiene alle due principali ss periodiche (α -elica, β sheet) o a nessuno delle due.

Usando poi finestre scorrevoli si può estendere la ricerca a tutta la sequenza

DeepMind di Google ha risolto un problema che affliggeva da 50 anni i biochimici: la ricostruzione 3D della forma che la proteina assume dopo il ripiegamento (folding) in base alla sua sequenza di aminoacidi.

Significa ad esempio accelerare la creazione di farmaci e vaccini, lo studio di alcune malattie.

La ricerca di nuovi farmaci e vaccini può essere accelerata grazie a quello che secondo molti esperti è uno dei principali risultati in biochimica del **2020** e una delle principali innovazioni in **intelligenza artificiale**:

il software **AlphaFold 2**

Quali metodi di predizione utilizzare?

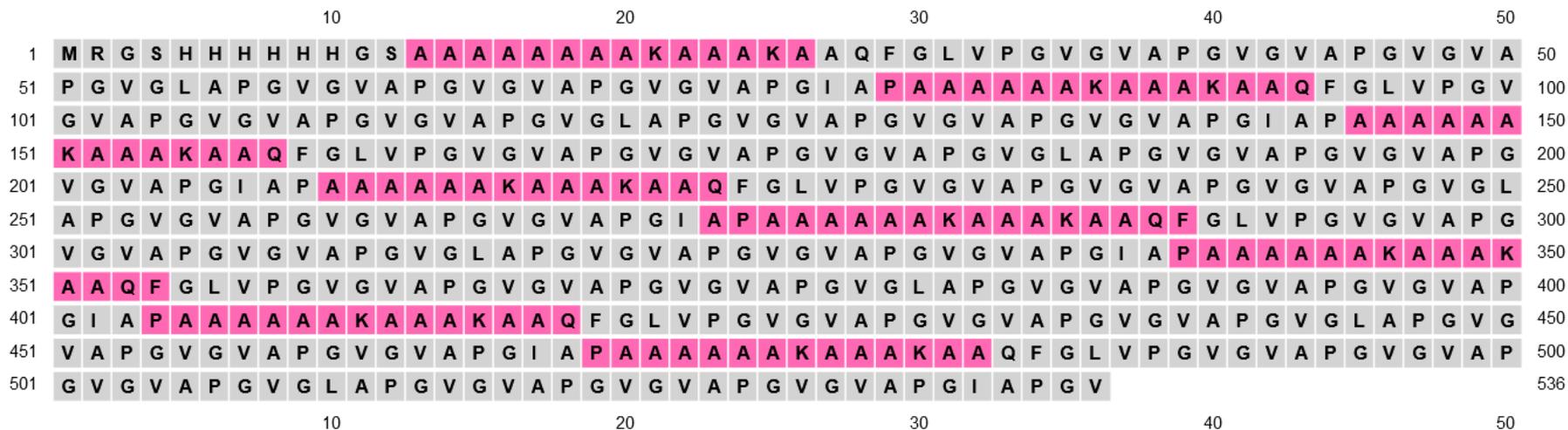
Neural network method	Accuracy (Q3)
Qian & Sejnowski 1988 (Qian and Sejnowski, 1988)	64.3%
PHD 1994 (Rost et al., 1994)	71.4%
PSIPRED 1997 (Jones, 1999)	76.5%
JPRED3 2008 (Cole et al., 2007)	81.5%
SPIDER3 2017 (Heffernan et al., 2017)	84%

Proteina HELP

PSIPRED

UCL Department of Computer Science: Bioinformatics Group

<http://bioinf.cs.ucl.ac.uk/psipred/>



Strand

Disordered, protein binding

Extracellular

Helix

Putative Domain Boundary

Re-entrant Helix

Coil

Membrane Interaction

Cytoplasmic

Disordered

Transmembrane Helix

Signal Peptide

Get PNG

Get SVG

Metodi Comparativi

Con il procedere dei progetti di sequenziamento stiamo assistendo ad una crescita esponenziale nel numero delle sequenze ma con poca conoscenza circa la loro struttura e funzione.

La determinazione della struttura e funzione di una sequenza non è un compito facile quindi la via migliore per arrivare alla comprensione della funzione e struttura delle sequenze è quella di mettere in relazione queste sequenze con altre sequenze conosciute **utilizzando metodi comparativi**.

I metodi di confronto delle sequenze rappresentano quindi il primo passo verso la caratterizzazione funzionale delle sequenze il cui compito è determinare **la relazione esistente tra struttura e funzione**.

I **metodi comparativi di analisi** possono essere applicati a diversi livelli:

- sequenza primaria
- struttura 3D

Le **metodologie più utilizzate nei metodi di analisi comparativi** sono:

- allineamento
- ricerca in banca dati

ALLINEAMENTO TRA SEQUENZE

L'allineamento tra due sequenze è fondamentale per determinarne la similarità reciproca ed inferire rapporti di omologia, funzione e conformazione strutturale.

Per raggiungere tale obiettivo occorrono due cose fondamentali:

- 1) Un algoritmo efficiente che sia in grado di rappresentare il più correttamente possibile la reale similarità tra le sequenze
- 2) Dei criteri di punteggio di similarità che siano in grado di definire la bontà dell'allineamento finale che si ottiene tra le sequenze e che supporti la creazione dell'allineamento stesso in fase di costruzione

Algoritmi euristici di allineamento

Sono nati insieme alle banche dati, con lo scopo di permettere una ricerca rapida, anche se meno accurata, utilizzando la similarità a coppie (o "pairwise") sulle migliaia di sequenze depositate.

Attualmente i programmi più utilizzati sono:

FASTA: Lipman & Pearson (1985)

Nasce con il nome di FAST-P (ricerca veloce di proteine), ma poi viene adattato per gli acidi nucleici (FAST-N), quindi riunito in un FAST per tutti, FAST-A, o semplicemente FASTA.

<http://www.ebi.ac.uk/Tools/fasta33/index.html>

BLAST: Altshul (1990)

Il Basic Local Analysis Search Tool viene ,

<http://blast.ncbi.nlm.nih.gov/Blast.cgi>

qualsiasi procedimento non rigoroso (a carattere approssimativo, intuitivo, analogico, ecc.) che consente di prevedere o rendere plausibile un risultato, il quale in un secondo tempo dovrà essere controllato e convalidato per via rigorosa.

FASTA

Protein | Nucleotide | Genomes | Proteomes | Whole Genome Shotgun | Web services | Help & Documentation | Also in this section | Feedback

Tools > Sequence Similarity Searching > FASTA

Protein Similarity Search

This tool provides sequence similarity searching against protein databases using the FASTA suite of programs. FASTA provides a heuristic query. FASTX and FASTY translate a DNA query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLOCAL (local database).

Web BLAST

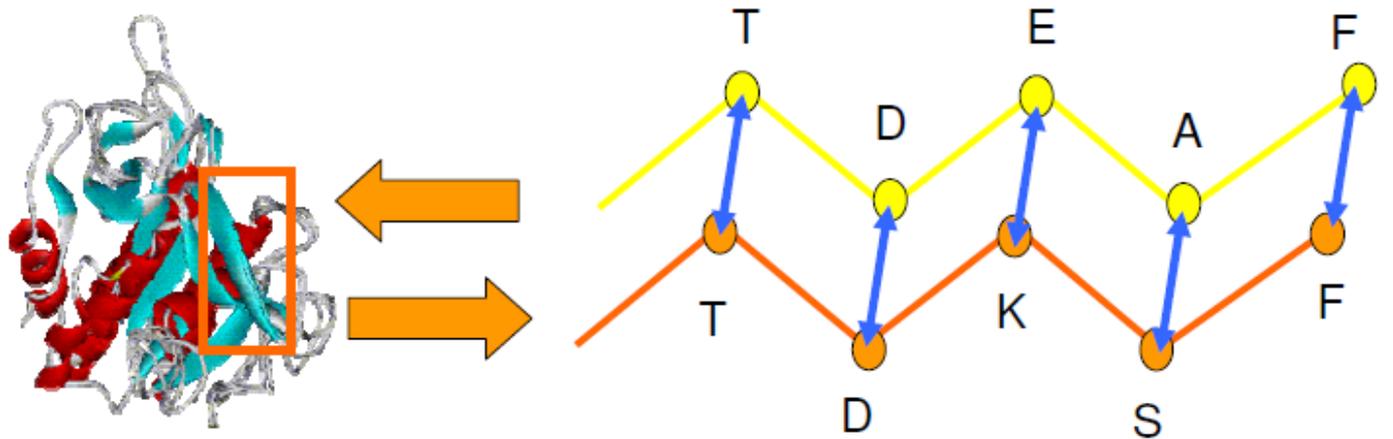
Nucleotide BLAST
nucleotide ► nucleotide

blastx
translated nucleotide ► protein

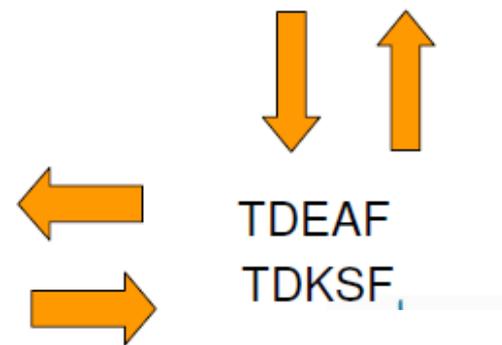
tblastn
protein ► translated nucleotide

Protein BLAST
protein ► protein

L'obiettivo finale: l'allineamento strutturale



MCDQTKHSKCCPAK---GNQCCPP--**TDEAF**---QQNQCCQSKGNQCCPPKQNQCCQPKG--
 M D +K ++CCP CCPP **TD F** Q++ CC + CCPPK + CC PK
 MSDSSKTNQCCPTPCCPPKPCPPKP**TDKSF**CCLQKSPCCPK--SPCCPPK-SPCCTPKVCP



TDEAF
 TDKSF

Allineare due sequenze (proteine o acidi nucleici)

- Cosa vuol dire allineare due sequenze? scrivere due sequenze orizzontalmente in modo da avere il maggior numero di simboli identici o simili in registro verticale anche introducendo intervalli (gaps – inserzioni/delezioni – *indels*)

seq1: TCATG

seq2: CATTG

TCAT-G
.CATTG

→ 4 caratteri uguali
1 inserzione/delezione

Altro problema è il punteggio che si deve dare all'allineamento ovvero la significatività

In definitiva i problemi sono:

Biologici

Trovare un modo per confrontare e rappresentare la similarità o “dissimilarità” tra le biomolecole

Computazionali

Trovare un modo per ottenere un match tra le sequenze di caratteri che compongono le biomolecole

Statistici

Come valutare la validità del risultato

Tre tipi di allineamento

- **globale**: vengono allineate le intere sequenze dall'N- al C-terminale (o dal 5' al 3' per il DNA)
- **locale**: vengono allineate solo le zone con la più alta densità di somiglianza generando uno o più suballineamenti
- **freeshift**: Una via di mezzo tra globale e locale

GAIOALE

D	L	G	P	S	S	K	Q	T	G	K	G	S	S	M	D	I	W	D	N	G	M
D	-	I	-	-	T	K	S	A	G	K	G	A	I	M	R	L	-	-	E	-	M

LOCALE

-	-	-	-	-	-	-	-	-	G	K	G	-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	G	K	G	-	-	-	-	-	-	-	-	-	-

FREESHIFT

D	L	G	P	S	S	K	Q	T	G	K	G	S	S	M	D	I	W	D	N	G	M
-	-	-	D	I	T	K	S	A	G	K	G	A	I	M	R	L	E	M	-	-	-

Predire la struttura terziaria

E di gran lunga la predizione più complessa che si possa fare su una proteina.

Esistono 3 metodi principali di predizione:

1 - Homology modelling:

se si conoscono proteine simili con struttura nota

2 – Fold recognition:

se si va alla ricerca di strutture simili, cercando il folding migliore

3 - Metodi di meccanica e dinamica molecolare

simulazioni di ripiegamento *in silico* molto complesse dato che si valutano tutte le possibili interazioni di tutti gli atomi con il solvente

Homology/
Comparative
modelling

Metodi di
meccanica e
dinamica
molecolare

Threading/
Fold
recognition

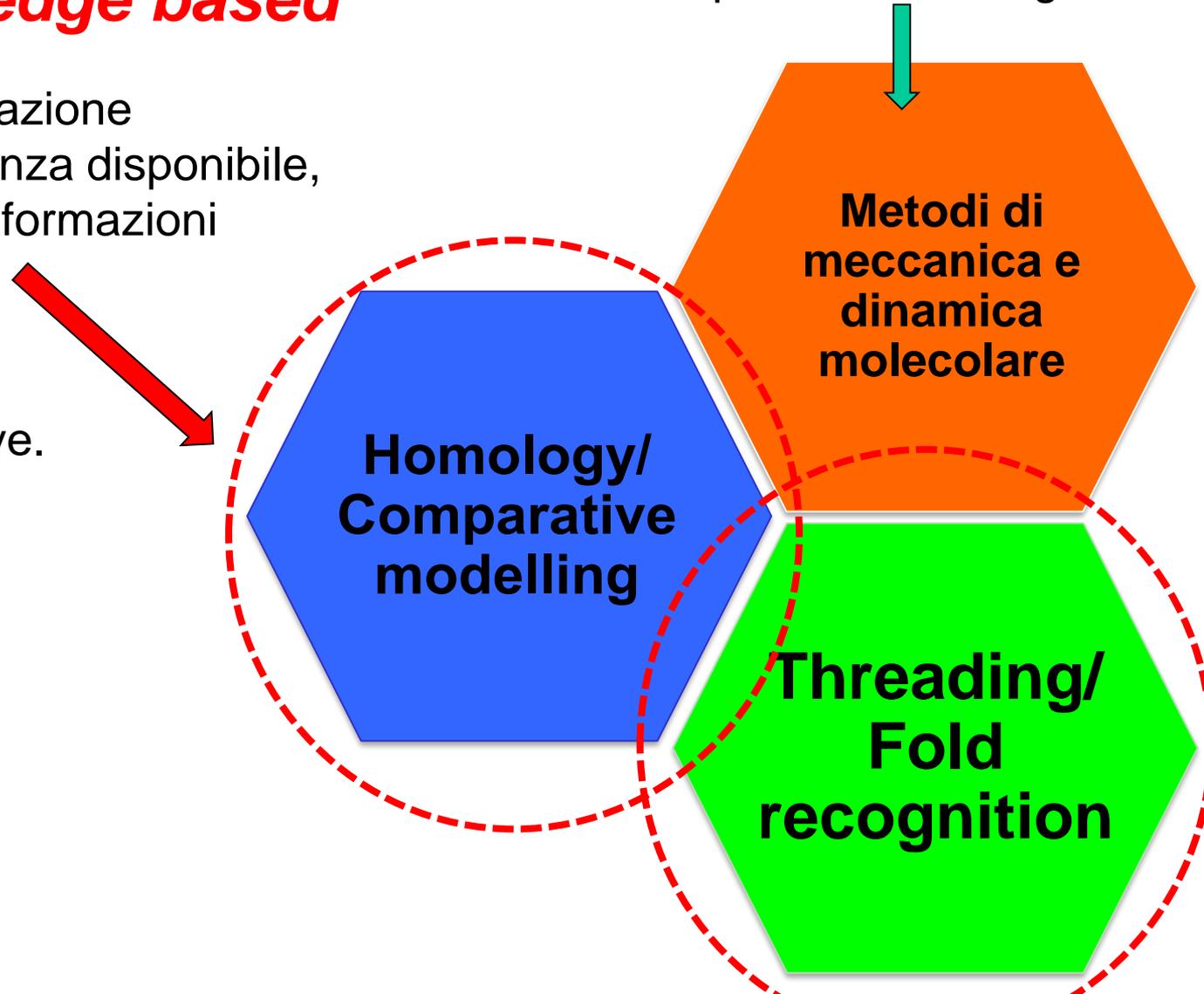
Metodi *knowledge based*

Si basano sull'informazione strutturale e di sequenza disponibile, utilizzando o meno informazioni evolutive.

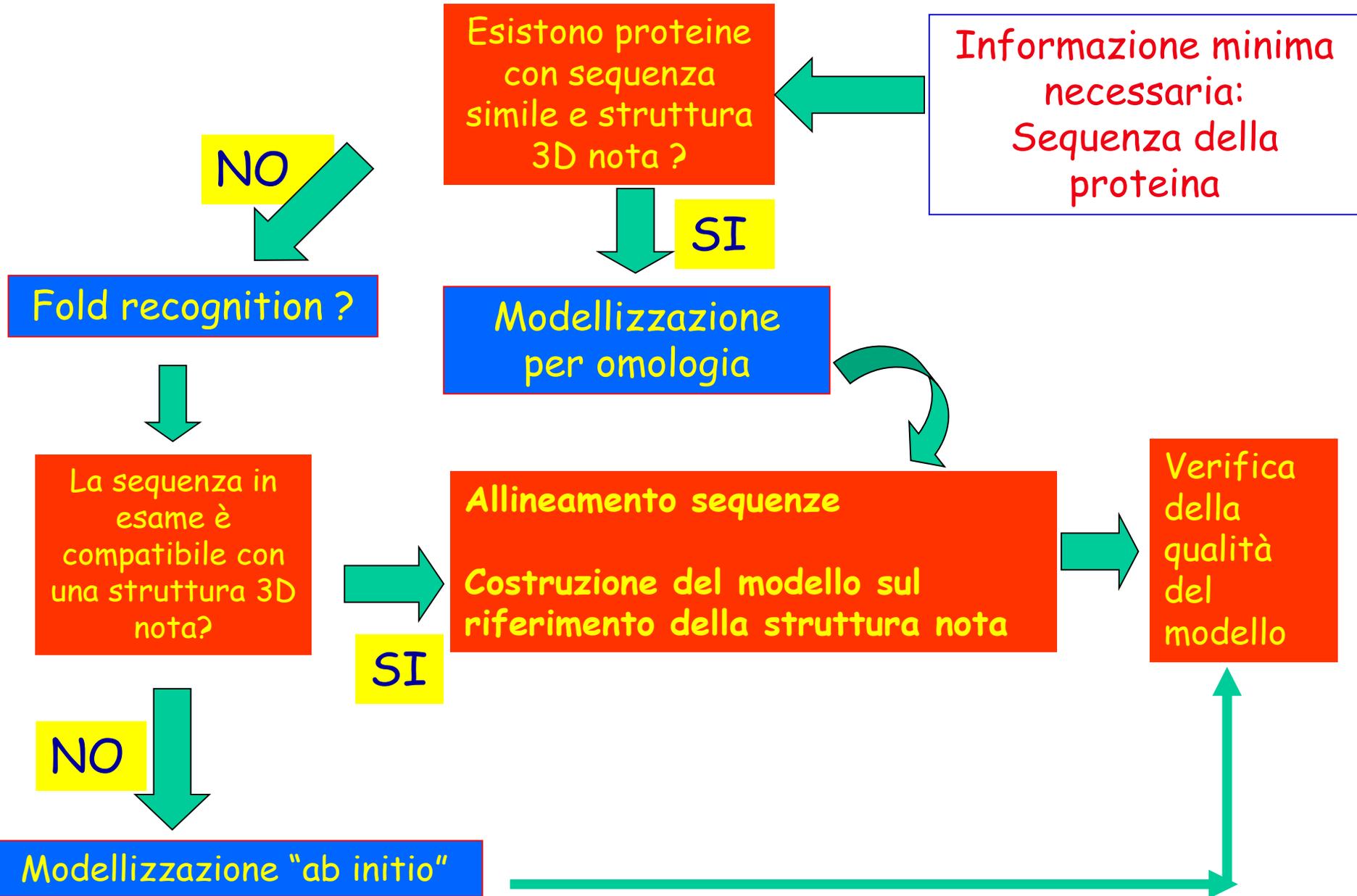
Possono dare ottimi risultati in tempo breve.

Si basano su principi teorici

tempi di calcolo lunghi !



Predizione della struttura tridimensionale delle proteine



FILOGENESI MOLECOLARE

Per filogenesi molecolare si intendono tutte quelle trasformazioni ed eventi di mutazione a livello del DNA che contribuiscono alla variabilità genetica e quindi ai suoi prodotti tra cui le proteine.

Il genoma subisce costantemente mutazioni ed i suoi meccanismi di riparo tengono sotto controllo la situazione affinché il tutto non degeneri in un accumulo dannoso per l'organismo in questione e la sua progenie (malattie genetiche trasmissibili, tumori ecc.)

Alle volte tali mutazioni sono eventi positivi che portano ad un vantaggio evolutivo premiato dalla selezione naturale che "fissa" tali mutazione nel DNA che viene trasmesso alle generazioni successive (eventi misurabili in milioni di anni)

Il DNA è un laboratorio in continuo sviluppo in cui intervengono fattori di selezione, ambientali ecc. Il tutto si traduce, per quanto ci riguarda, in nuove proteine che si modificano, acquisiscono nuove funzioni.

Le variazioni genetiche sono il presupposto fondamentale per l'evoluzione biologica: hanno origine spontaneamente in seguito a errori nel processo della replicazione oppure a mutazioni accidentali, anche dovute a fattori ambientali, che in qualche modo alterano la sequenza del DNA.

Definizioni: l'identità, la similitudine, la conservazione

Identità

La misura in cui due sequenze (di nucleotidi o aminoacidi) sono invarianti. (es. identità del 32% => 32 a.a. su 100 sono ordinatamente identici)

Conservazione

In una sequenza, modifiche in una specifica posizione di un amminoacido (o meno comunemente, DNA) che preservano le proprietà fisico-chimiche del residuo originale.

Similitudine

La misura in cui due sequenze (di nucleotidi o aminoacidi) sono correlate. Si basa su identità + conservazione.

Similarita' e omologia

- Due sequenze sono simili se possono essere allineate in modo che molti ammino acidi corrispondenti sono identici o simili
- Tecnicamente due o piu' sequenze possono essere definite omologhe se derivano da un progenitore comune
- L'omologia tra due sequenze si deduce dalla loro similarita in sequenza o funzione.

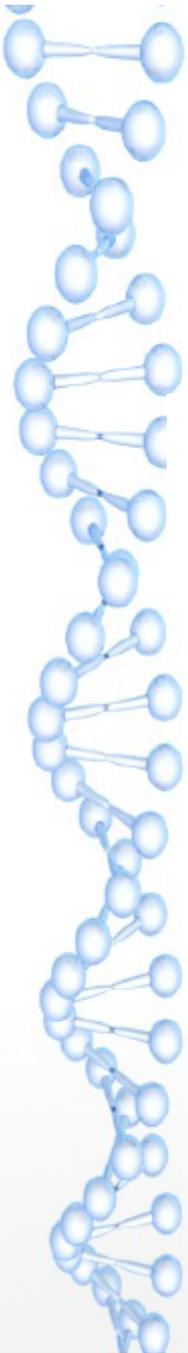
Definizione: omologia

Omologia

Similitudine attribuita a discendenti da un antenato comune.

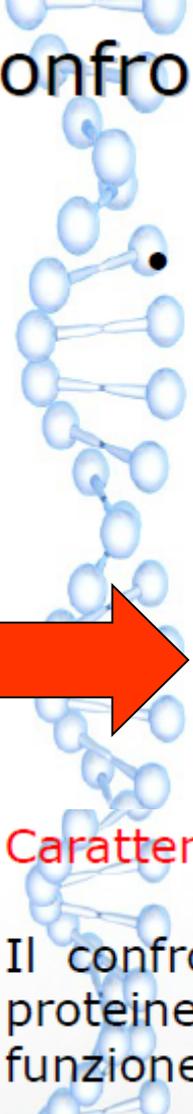
NOTA BENE:

- **OMOLOGIA** indica che due entità (es. 2 sequenze) hanno una stessa origine filogenetica, cioè derivano da un antenato comune. È un carattere **QUALITATIVO**.
- **SIMILITUDINE** indica che due entità (es. 2 sequenze), in relazione ad *un certo criterio* comparativo, hanno *un certo grado* di similitudine. È un carattere **QUANTITATIVO**



Similarità e omologia

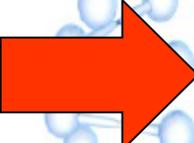
- Tra due o più sequenze può esserci un certo grado di similarità.
- A volte una similarità tra sequenze implica una similarità strutturale e, conseguentemente, una similarità funzionale.
- L'omologia tra sequenze indica invece una comune origine evolutiva tra di esse. Due sequenze si dicono omologhe quando discendono entrambe da una sequenza ancestrale comune(ancestore)
- Due o più sequenze simili tra loro possono quindi essere omologhe o meno.



Confrontare sequenze

- Il confronto fra sequenze, nucleotidiche o aminoacidiche, è uno dei compiti fondamentali della bioinformatica.

Perché è possibile confrontare sequenze?



Perché generalmente in natura le strutture molecolari non vengono create ex-novo ma per modificazione di modelli preesistenti.

Caratterizzazione di proteine con funzione ignota

- Il confronto di una proteina a funzione ignota con una famiglia di proteine a funzione nota può permettere di formulare ipotesi sulla funzione della prima.

Relazione sequenza-struttura-funzione

Evoluzione divergente

- proteine con elevata identità di sequenza discendono da un ancestore comune
- avranno strutture molto simili

casi limite

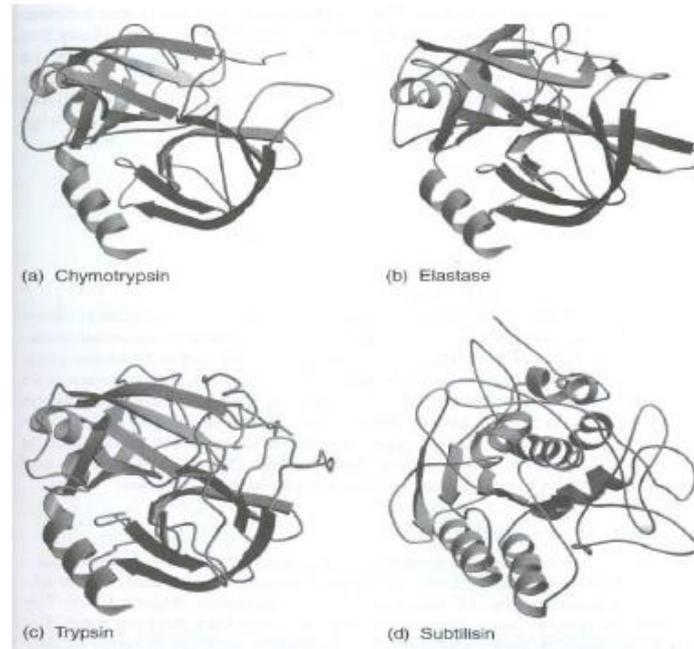
- 1) proteine con bassa similarità di sequenza ma struttura complessiva e di sito attivo simili sono probabilmente omologhe
 - ⇒ conservazione selettiva di residui strutturalmente e funzionalmente importanti
- 2) strutture e sequenze simili ma funzioni biochimiche differenti
 - ⇒ le strutture divergono più lentamente delle funzioni durante l'evoluzione

Evoluzione convergente

- proteine che differiscono in sequenza e struttura
- presentano però una configurazione del sito attivo che converge verso una medesima struttura, che ne determina la funzione biochimica

Evoluzione divergente di famiglie di proteine

Le serino-proteasi sono una famiglia di enzimi che hanno una struttura 3D molto simile e un pressoché identico meccanismo di reazione.



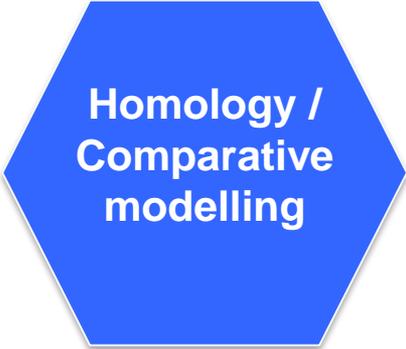
La differenza fra questi enzimi è la **specificità di substrato**:

- La tripsina taglia i peptidi in corrispondenza di Lys e Arg
- La chimotripsina taglia i peptidi in corrispondenza di Trp, Phe, Tyr
- l'elastasi taglia i peptidi in corrispondenza di piccoli Aa idrofobici (Ala)

evoluzione convergente

la serino-proteasi del batterio è diversa come struttura 3D dalle serino-proteasi di mammifero ma ha la stessa specificità di substrato e lo stesso meccanismo catalitico.

Organismi diversi partendo da una differente struttura 3D hanno evoluto proteine con una stessa funzione.



Homology /
Comparative
modelling

Modelling per Omologia

(Homology (o Comparative) Modelling)

- In generale, a maggiore identità di sequenza tra due proteine, corrisponde maggiore similarità tra strutture
- La qualità del modello dipende dalla similarità tra le sequenze delle due proteine

Homology modelling

Assunzione di base: due proteine che presentano una identità del 30% circa, molto probabilmente avranno una struttura simile.

Si va a valutare la

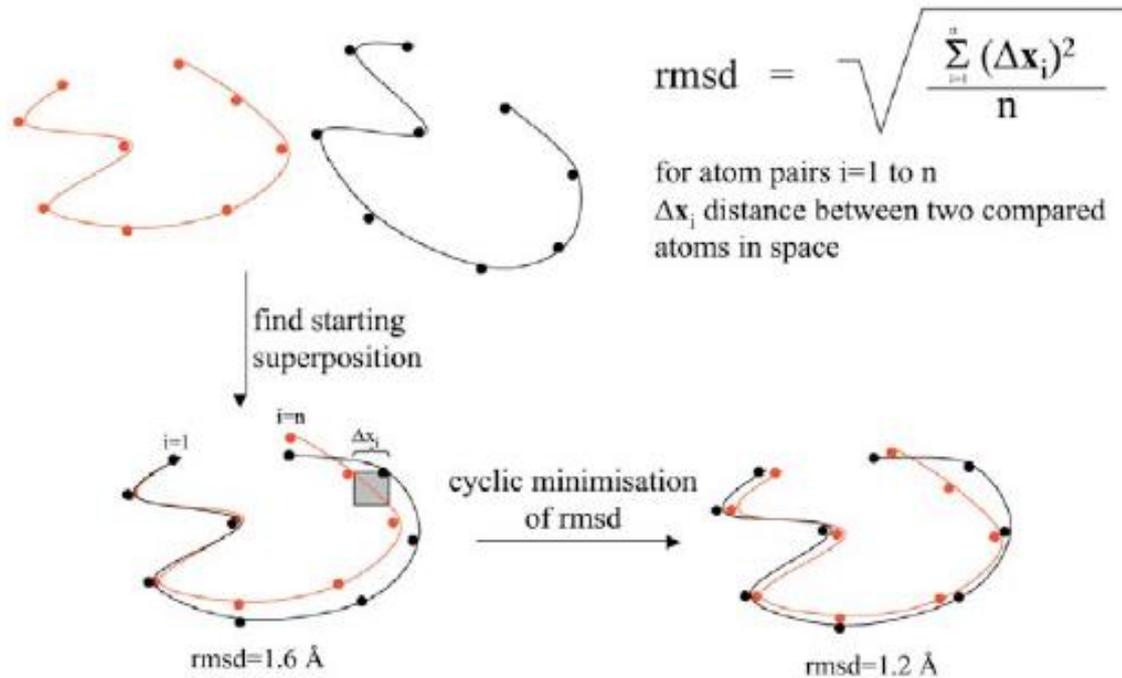
r.m.s.d. (root mean square deviation) =
$$\sqrt{\frac{\sum_{i=1}^n (\Delta x_i)^2}{n}}$$

cioè la distanza quadratica media tra gli atomi, generalmente i carboni alpha, ma si può calcolare anche su tutti gli atomi.

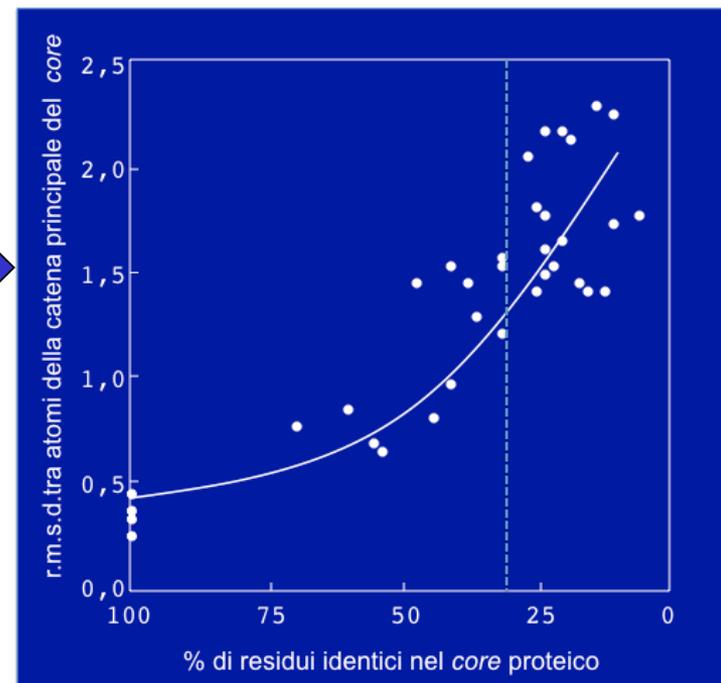
Minore è il r.m.s.d. maggiore è la similarità strutturale

=> è stato osservato che proteine che condividono un'identità di sequenza del 50% mantengono circa il 90% dei residui in posizioni conservate

Root mean square deviation (rmsd)



Se l'identità tra due sequenze proteiche è **superiore** al 30%, si può assumere che le loro strutture siano simili



Allineamento di Sequenze

- ▶ L'allineamento tra due o più sequenze può aiutare a trovare regioni simili per le quali si può supporre svolgano la stessa funzione;
- ▶ La similarità tra due o più sequenze può essere definita in base a una funzione distanza: Tanto più simili sono le sequenze, tanto meno distanti sono;
- ▶ Esistono diversi algoritmi di allineamento ciascuno dei quali definisce una funzione distanza;
- ▶ Dato un allineamento possiamo assegnare uno **Score** che indica il grado di similarità delle due sequenze.

Quindi...

- Proteine con sequenze simili hanno generalmente anche strutture tridimensionali simili
- Proteine simili da un punto di vista strutturale tendono ad avere anche funzioni simili (non sempre)
- Proteine con funzione simile potrebbero avere una sequenza completamente differente (il sito attivo ha una struttura simile derivante da evoluzione convergente)

L'approccio all'homology modeling è abbastanza intuitivo, ma è bene seguire delle considerazioni pratiche che possono essere definite

Linee guida per costruire un buon modello:

- ① Analizzare la struttura secondaria prima di quella tridimensionale, dato che se ci sono regioni fortemente random-coil, è bene escluderle dalla modellazione (sono, per definizione, non modellabili).
- ② Identificare in banca dati una proteina a struttura nota che abbia una alta identità (> 50%) con la propria, o identità globale o identità locale.
- ③ Allineare al meglio le due proteine. Rifinire l'allineamento a mano, se necessario. Questa tappa è critica, perchè la modellazione verrà fatta sulle proteine allineate.

Homology/ Comparative modelling

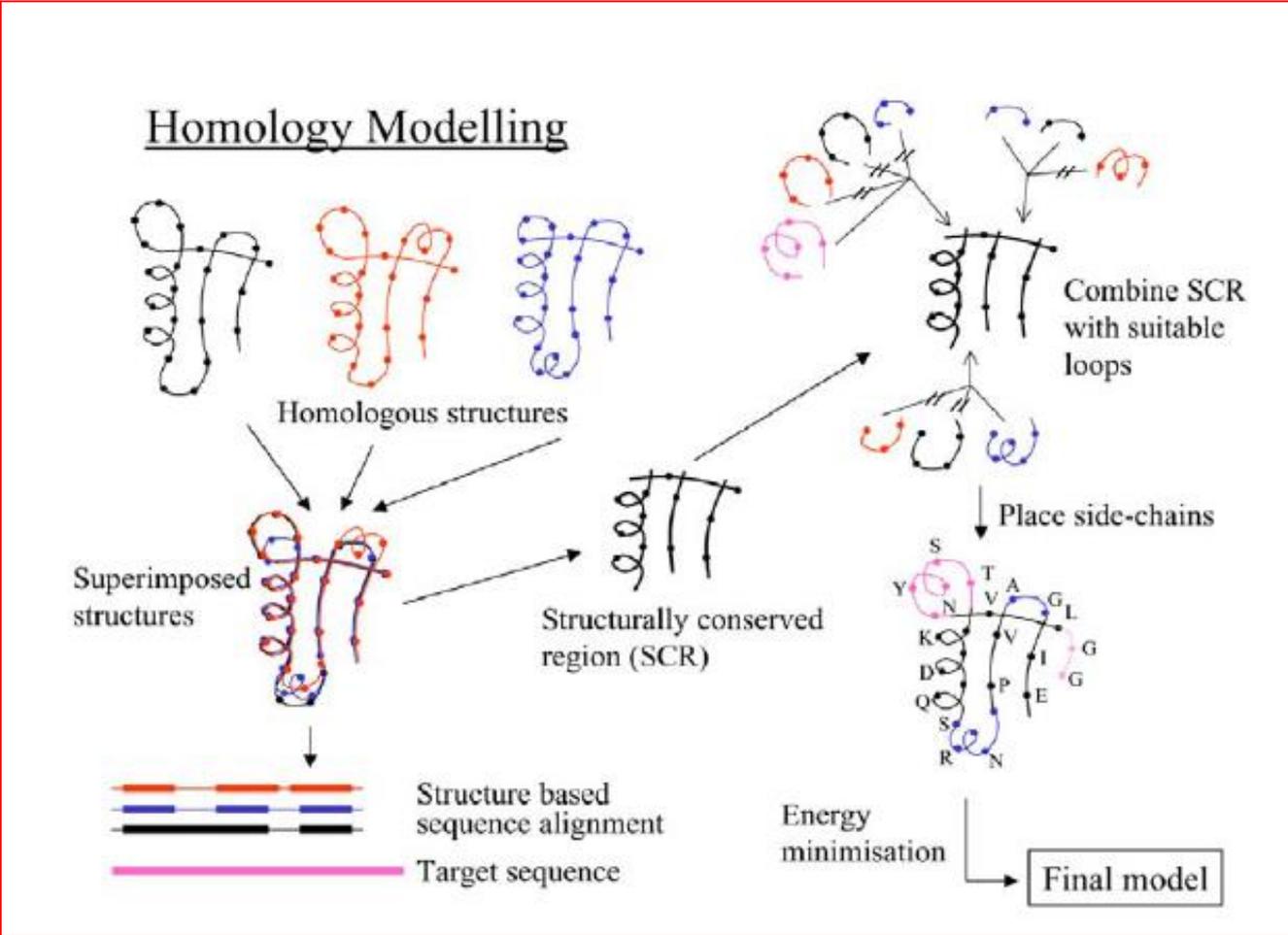
4. Cercare in banca dati il più alto numero di proteine simili tra loro e identificare dopo il multiallineamento le regioni di struttura secondaria conservate (in genere se ci sono dal multiallineamento si vede). Si costruisce così un *pre-modello*.
5. Modellare per prima cosa le regioni altamente variabili (in genere i loops) che connettono regioni a struttura secondaria definita. Esistono database anche dei loops (tutti i loops osservati), quindi non si modella a caso, ma si cerca la struttura di un loop già presente. Se non c'è, o se ce ne sono più di uno compatibile, osservare le regioni pre-loop e post-loop.
6. Modellare le catene laterali sulla base delle catene della proteina nota. Esistono per questo delle librerie di ROTAMERI, cioè isomeri delle catene laterali con angoli di torsione adeguati alle torsioni del backbone.

Homology/ Comparative modelling

7 - Risolvere, se possibile, i problemi relativi alle collisioni dei vari atomi, o manualmente o con programmi che indicano l'energia minima delle strutture. Alla fine gli atomi di una proteina non dovranno collidere, e la proteina dovrà essere alla minor energia possibile.

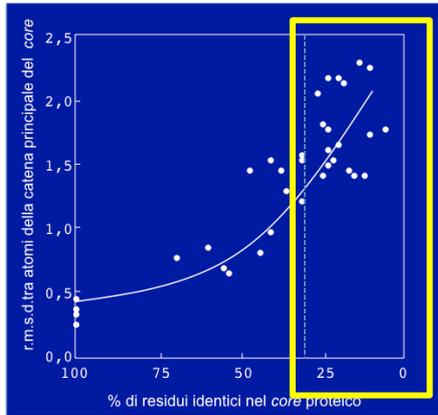
**Nonostante ci siano delle regole abbastanza precise,
l'homology modelling è molto complesso e richiede
un grande lavoro e molto tempo.**

schematizzando....



Protein threading = fold recognition

Threading /
Fold
recognition



Al di sotto del 20-25% di identità tra *query* e *templato* (struttura risolta) ci troviamo nella cosiddetta “zona notturna” (*midnight zone*)

in cui non si riesce a distinguere una somiglianza di sequenza significativa, ed anzi le similarità possono derivare sia da evoluzione divergente (omologia) che convergente.

In questo caso si utilizza un approccio diverso, detto di **fold recognition**, per mezzo del **threading** (allineamento).

Protein threading = fold recognition

Il *Protein threading* si basa su due osservazioni fondamentali:

- 1) le proteine presenti in natura hanno un numero finito di strutture possibili, o almeno un **numero finito di topologie (<1000)**
- 2) Il **90%** delle strutture nuove archiviate nel data-base PDB negli ultimi 3 anni hanno struttura simile ad una già esistente nel PDB.



Questo tipo di approccio parte dall'osservazione che la struttura tridimensionale è generalmente più conservata della sequenza: si stima che infatti possano esistere 10^4 tipi di *fold* base, a fronte di un numero virtualmente infinito di possibili sequenze diverse.

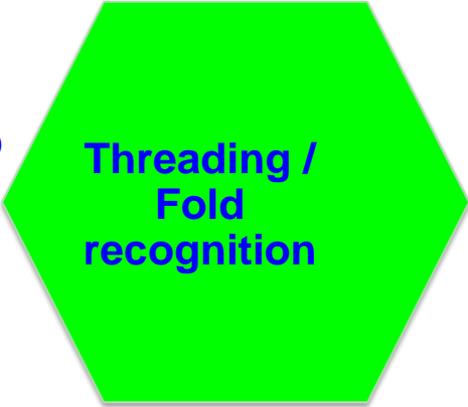
Il metodo è usato per predire la struttura di quei biopolimeri che hanno gli stessi ripiegamenti (*fold*s) di biopolimeri di struttura nota ma non presentano omologia con tali strutture note.

La predizione viene effettuata per "*threading*" (cioè allineamento) di ogni aminoacido nella posizione della sequenza della struttura nota (templato) e valutando quanto buono risulta il fitting.

Un ottimo server per predizioni di questo tipo è **Phyre** (<http://www.sbg.bio.ic.ac.uk/phyre2/>).

Threading /
Fold
recognition

Tra i diversi sistemi di riconoscimento del *fold*, uno dei primi proposti è quello dei **Profili-3D**, in cui la struttura 3D viene “tradotta” in una sequenza che tiene conto delle caratteristiche strutturali di ogni residuo.



Threading /
Fold
recognition

In particolare, i residui vengono divisi in **3** diverse classi:

- **esposto al solvente,**
- **parzialmente esposto,**
- **non esposto**

queste classi sono poi suddivise in **sottoclassi**, per un totale di **6**.

Il residuo può poi appartenere a strutture secondarie (eliche e foglietti) o *loop*.

In totale, esistono **6x3=18 classi**, che descrivono l'intera gamma dei possibili intorni chimici dei residui di una struttura.

JMB



A 3D-1D Substitution Matrix for Protein Fold Recognition that Includes Predicted Secondary Structure of the Sequence

Danny W. Rice and David Eisenberg*

Or

Threading /
Fold
recognition

Per mezzo di matrici di punteggio 3D-1D,

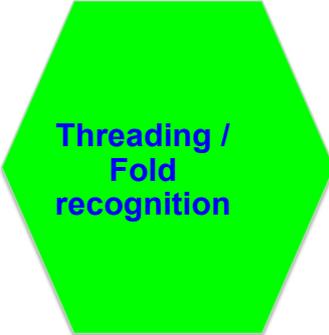
The 1D-sequence/1D-string scoring matrix

ottenute analizzando strutture reali, è possibile confrontare questo alfabeto di 18 lettere con quello a 20 lettere degli aminoacidi: in pratica, queste matrici ci dicono con quale propensione ogni aminoacido si colloca in ognuna delle 18 classi.

E' possibile così confrontare una sequenza proteica allineandola con tutti i possibili profili di tutti i fold conosciuti, ricavando un punteggio che valuti il *best fitting* della sequenza.

Facendo così si è in grado di identificare strutture di proteine anche molto divergenti tra loro, al punto di non essere riconosciute da nessun programma di allineamento o di similarity search.

Un esempio tipico è l'individuazione della struttura di proteine che hanno la stessa funzione a causa di una **evoluzione convergente**: originandosi da geni diversi non correlati, la sequenza (sia aminoacidica, sia nucleotidica) sarà molto diversa, ma la struttura terziaria, almeno nell'intorno del sito catalitico, deve essere costante per garantire una stessa funzionalità



Threading /
Fold
recognition

Iterative Threading ASSEmbly Refinement

University of Michigan

Department of Computational Medicine and Bioinformatics



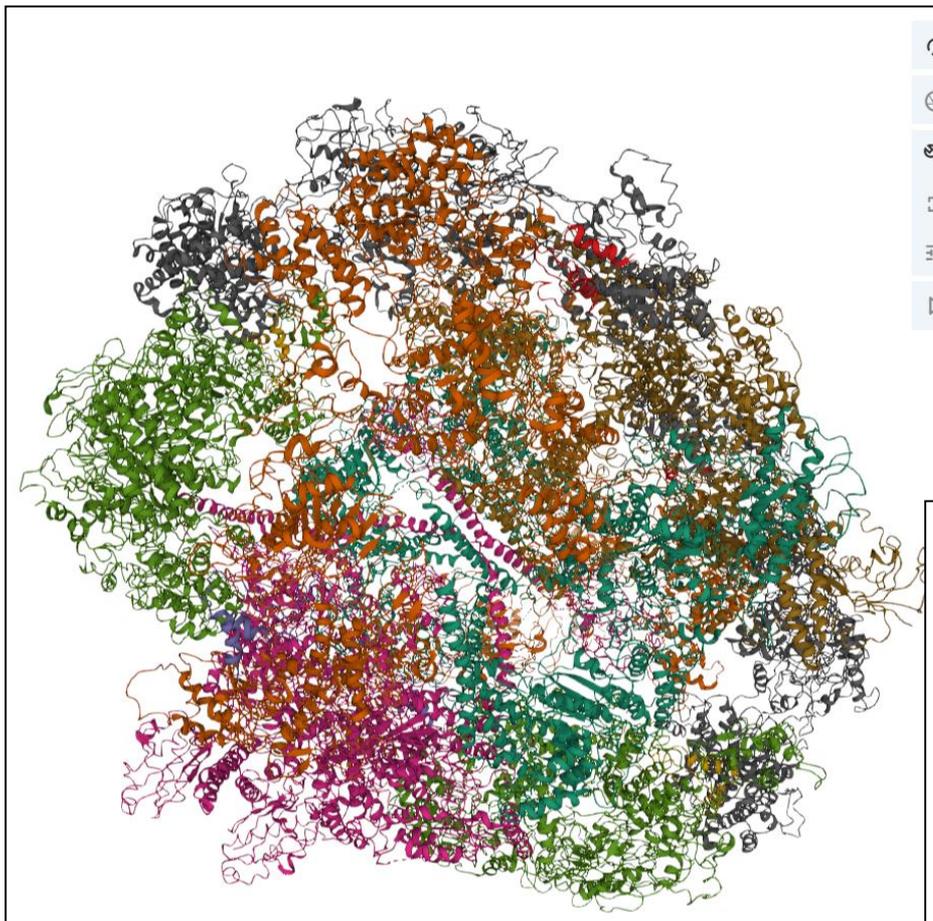
I-TASSER

Protein Structure & Function Predictions

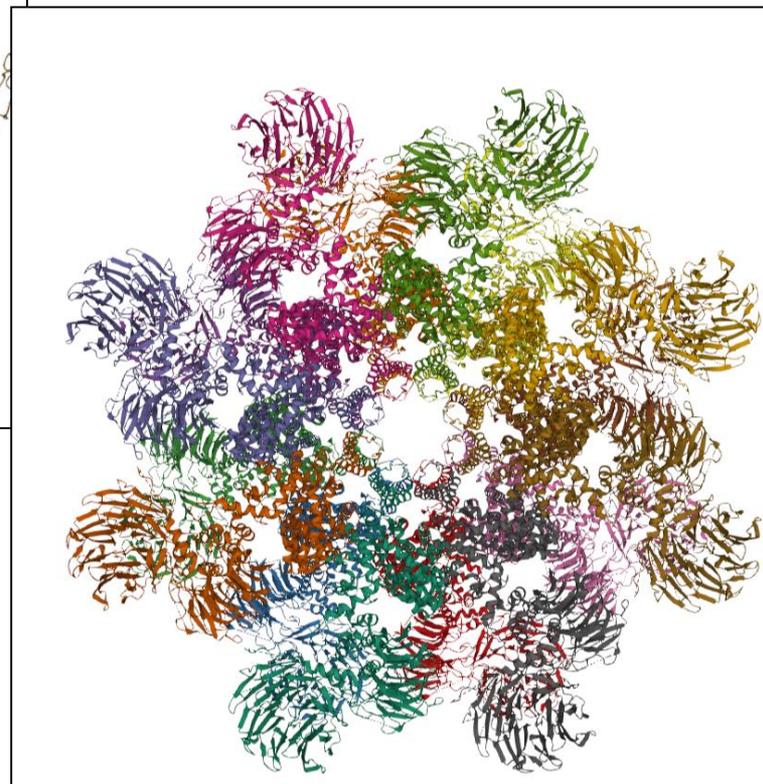
(The server completed predictions for [579090 proteins](#) submitted by [138555 users](#) from [150 countries or regions](#))

(The template library was updated on [2020/11/01](#))

I-TASSER (Iterative Threading ASSEmbly Refinement) is a hierarchical approach to protein structure prediction and structure-based function annotation. It first identifies structural templates from the PDB by multiple threading approach [LOMETS](#), with full-length atomic models constructed by iterative template-based fragment assembly simulations. Function insights of the target are then derived by re-threading the 3D models through protein function database [BioLiP](#). I-TASSER (as 'Zhang-Server') was ranked as the No 1 server for protein structure prediction in recent community-wide [CASP7](#), [CASP8](#), [CASP9](#), [CASP10](#), [CASP11](#), [CASP12](#), and [CASP13](#) experiments. It was also ranked the best for function prediction in [CASP9](#). The server is in active development with the goal to provide the most accurate protein structure and function predictions using state-of-the-art algorithms. Please report problems and questions at [I-TASSER message board](#) and our developers will study and answer the questions accordingly. ([>> More about the server ...](#))



2PFF
Structural Insights of Yeast Fatty Acid
Synthase
(rank=1)



4V4L (1VT4)
Structure of the Drosophila apoptosome (rank=3)

Protein threading VS. Homology

Quando l'identità di sequenza in un allineamento sequenza-sequenza è bassa (cioè <25%), *l'homology modelling* potrebbe non produrre una previsione significativa. In questo caso, il *protein threading* può generare una migliore previsione.

Quando non si trova un'omologia significativa, il *protein threading* può fare una previsione basata sulle informazioni della struttura.

La modellazione dell'omologia è per quei sistemi che hanno proteine omologhe con struttura nota (di solito della stessa famiglia), mentre il *protein threading* è per quei sistemi per i quali è stata trovata solo un'omologia a livello di *folds*.

riassumendo...

Homology modelling	Threading/Fold-recognition
Identifica prima gli omologhi	Prova tutte le possibili strutture
Si determina l'allineamento ottimale	Prova tutti i possibili allineamenti strutturali
Ottimizza un modello	Valuta molti modelli poco accurati nei dettagli

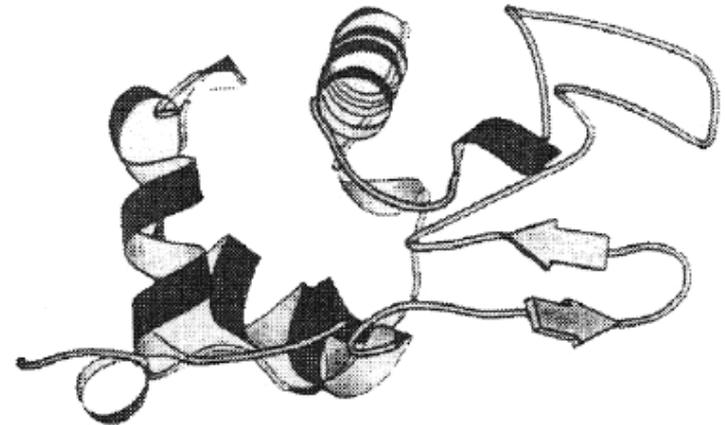
ESEMPIO:

OMOLOGIA E FUNZIONE

può non essere sufficiente a determinare la categoria funzionale anche nel caso di omologia di sequenza confermata ed appartenenza ad una stessa famiglia di proteine sebbene questo caso sia più raro.



(a) Lysozyme EC 3.2.1.17

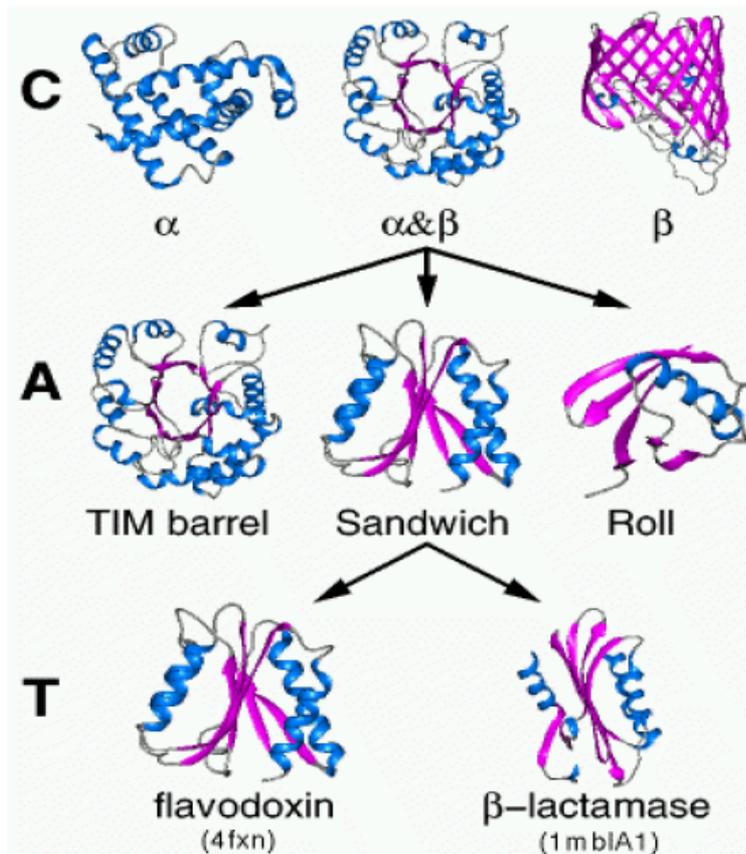


(b) Alpha-lactalbumin (nonenzyme)

Lisozima e alpha-lactalbumina hanno il 40% di identità e struttura simile ma differente funzione. Il sugar-binding site delle due proteine è stato mantenuto ma nella lactalbumina ha cessato la sua funzione

Banche dati – Proteine: CATH

- Database che definiscono *famiglie strutturali*.
- Aiutano a predire le strutture e a caratterizzarle (secondo un'idea evuzionistica della funzione).



CATH: a hierarchical domain classification of protein structures in the Protein Data Bank

Classificazione in modo curato sulla base di:

- **CLASSE** (contenuto e tipo di strutture secondarie)
- **ARCHITETTURA** (descrizione dell'orientamento delle strutture secondarie senza tener conto delle connessioni)
- **TOPOLOGIA** (tiene conto delle connessioni che caratterizzano le strutture secondarie)
- **(H)OMOLOGIA** (raggruppa proteine con strutture e funzioni simili)

Che cos'è CATH-Gene3D?

CATH è una classificazione delle strutture proteiche scaricate dalla Protein Data Bank.

Raggruppano i domini proteici in superfamiglie quando esistono prove sufficienti della loro divergenza da un antenato comune.

- Search CATH by text, ID or keyword
- Search CATH by protein sequence
- Search CATH by PDB structure
- Browse CATH Hierarchy
- CATH Release Statistics
- CATH Tutorials

Gene3D utilizza le informazioni di CATH per prevedere la posizione dei domini strutturali su milioni di sequenze proteiche disponibili in database pubblici.

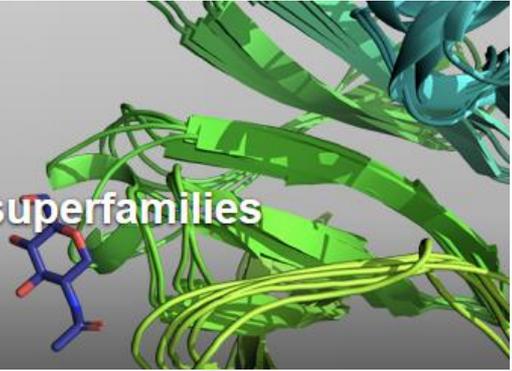
Questo ci permette di includere annotazioni aggiuntive nel database CATH-Gene3D, come informazioni funzionali e residui del sito attivo.

<https://www.cathdb.info/>

CATH / Gene3D v4.3

151 million protein domains classified into 5,481 superfamilies

Search by keywords, PDB code, GO term, etc



3D Structure

Find out what 3D structure your protein adopts

[Find out more](#)

[Go](#)



Protein Evolution

Learn about a particular protein family and how it evolved

[Find out more](#)



Protein Function

Investigate the function of your protein

[Find out more](#)

[Go](#)



Conserved Sites

Look at protein sites that are highly conserved and implicated in function

[Find out more](#)

[Go](#)



Download Data

Download data files and query CATH via webservice

[Go](#)



Learn more

Find out how CATH is created and maintained, how to link to CATH and more

[Go](#)