

# Struttura delle proteine

**Struttura primaria**

**Struttura secondaria** → **Dicroismo circolare**

**Metodi di predizione di  
struttura secondaria**

**Struttura terziaria**

→ **Cristallografia ai RX**

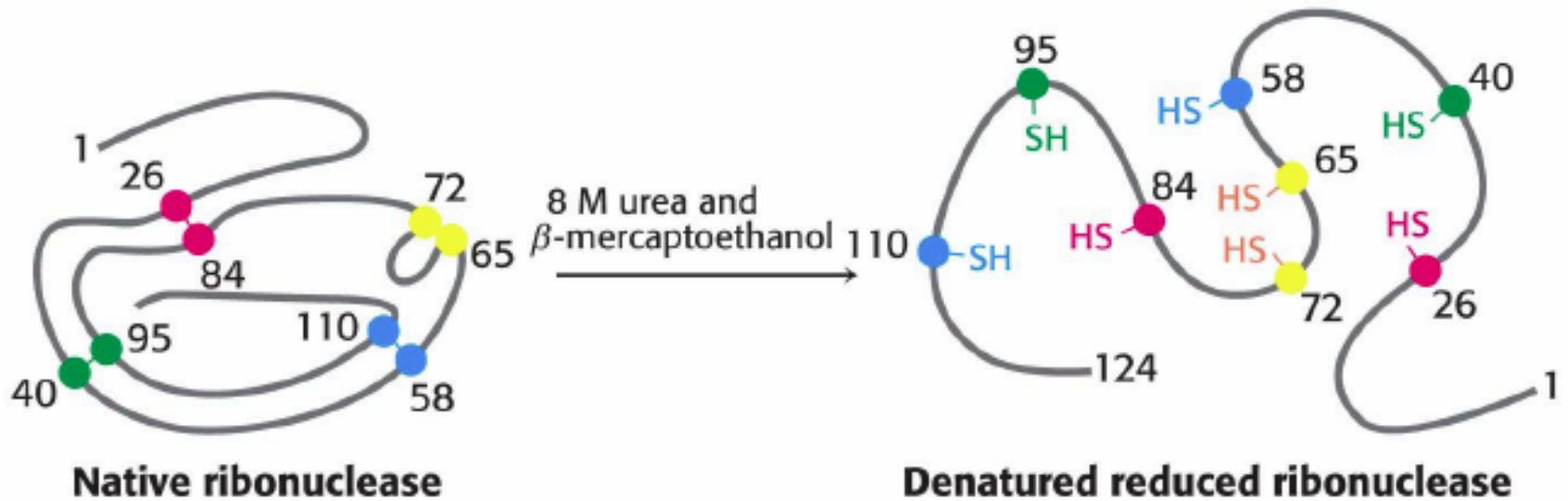
→ **NMR**

**Homology Modelling  
Fold Recognition  
Folding ab-initio**

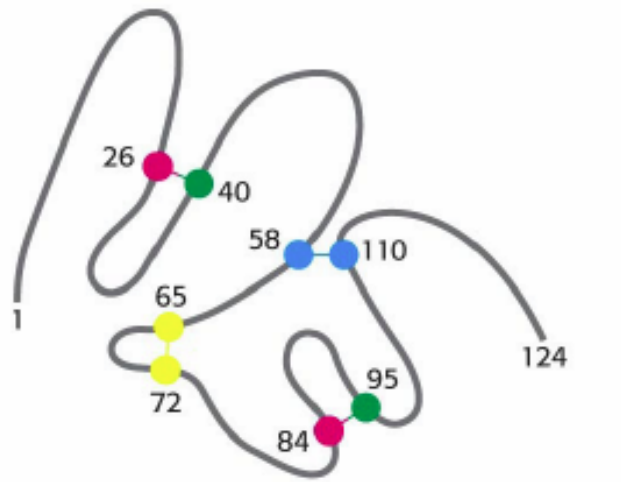
**Struttura quaternaria**

## Dalla struttura primaria alla terziaria ...

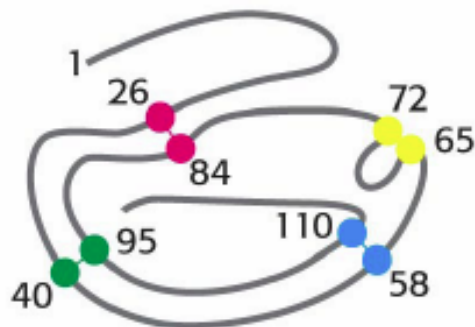
### Principio di Anfinsen



Denaturando la ribonucleasi, dalla conformazione nativa si ottiene uno stato “disordinato”. La denaturazione avviene per effetto dell’urea, mentre il beta-mercaptoetanolo riduce i ponti S-S



Trace of  
 $\beta$ -mercaptoethanol



**Native ribonuclease**

L'allontanamento completo dell'urea e del mercaptoetanolo provoca la formazione di ponti S-S casuali e la formazione di conformazioni diverse da quella nativa ("conformazioni arruffate")

Aggiungendo piccole quantità di mercaptoetanolo, si provoca la rottura dei ponti S-S casuali e la formazione di nuovi ponti S-S fino a che si raggiunge l'equilibrio finale, quando il ripiegamento nella conformazione più stabile permette la formazione dei legami S-S originali e la formazione della ribonucleasi nativa

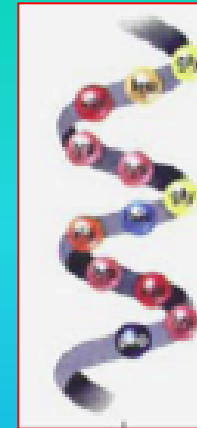
**L'organizzazione di struttura secondaria e terziaria di una proteina è determinata dalla struttura primaria.**

**Circa il 50% del backbone di una proteina si ripiega formando strutture secondarie come l'alfa elica o il foglietto beta.**

**E' più semplice predire la struttura ad alfa elica che non il foglietto beta:** infatti, l'alfa elica è determinata da interazioni locali (ovvero che si formano tra amminoacidi a poca distanza nella sequenza), mentre la struttura a foglietto beta-strand è molto più dipendente dal contesto: servono infatti due o più beta-strands affiancati per formare un foglietto beta, e gli strands possono essere anche molto lontani nella sequenza.

## Dalla sequenza alla struttura secondaria

Gly-Leu-Val-Lys-Lys-Gly-His-Ala-Lys-Val-Lys-Pro



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, ...
- **Consenso di metodi**: SOPM, SOPMA, JPRED, ...

## Predizione di strutture secondarie

### Metodo Chou-Fasman:

Sviluppato negli anni '70, si basa su una procedura statistica che valuta la propensione di ogni amminoacido di far parte di una struttura secondaria.

Ogni amminoacido viene classificato per la sua propensione ad entrare in strutture secondarie come “**former**”, “**breaker**” o “**indifferent**”.

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

Conformational Preferences of the Amino Acids

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

## Predizione di strutture secondarie

### Metodo Chou-Fasman:

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

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

Metodo GOR: Sviluppato negli anni '70, si basa su una procedura simile a quella del metodo Chou-Fasman, ma usa finestre di lunghezza maggiore.

## Caratteristiche chimico-fisiche e riconoscimento di pattern:

Metodi di predizione che si avvalgono del riconoscimento di pattern strutturali specifici o di caratteristiche chimico-fisiche per identificare la presenza di elementi di struttura secondaria.

Possono usare allineamenti multipli di sequenze anziché sequenze singole, e tengono conto di:

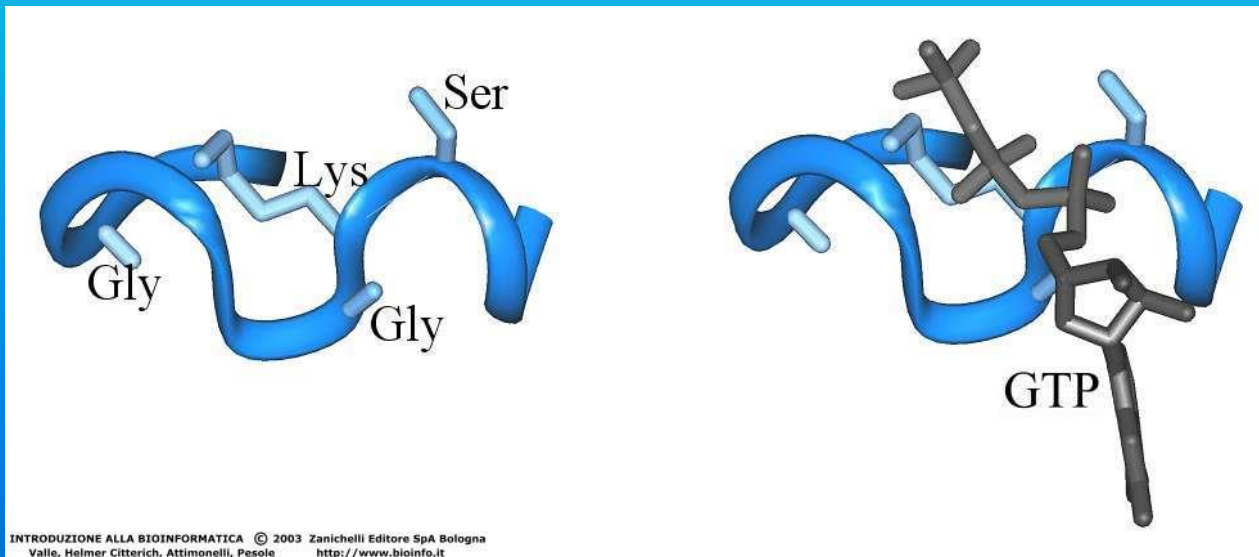
- **Posizioni di inserzioni e delezioni** (di solito in corrispondenza di loop)
- **Gly e Pro conservate** (presenza di beta turn)
- **Residui polari e idrofobici alternati** (presenza di beta strand di superficie)
- **Amminoacidi idrofobici e idrofili con periodicità 3.6** (alfa eliche anfifiliche)

La predittività con questi metodi migliora di circa 8-9% rispetto ai soli metodi statistici.



## Definizione di pattern

Un **pattern** è costituito da un insieme di caratteri (nucleotidi o amminoacidi) non necessariamente contigui nella sequenza ma che si trovano sempre o sono spesso associati ad una precisa struttura e funzione biologica (ad esempio: promotori o hanno la stessa capacità di legare nucleotidi)



## PHD

**Alla base del metodo c'è l'osservazione che in un allineamento multiplo si evidenziano conservazioni di amminoacidi che rispettano la conservazione della struttura.**

**La singola query viene confrontata con le sequenze presenti in banche dati per trovare proteine simili. La query e le proteine simili vengono allineate tutte insieme.**

**Quindi, l'allineamento multiplo è usato come input della rete neurale.**

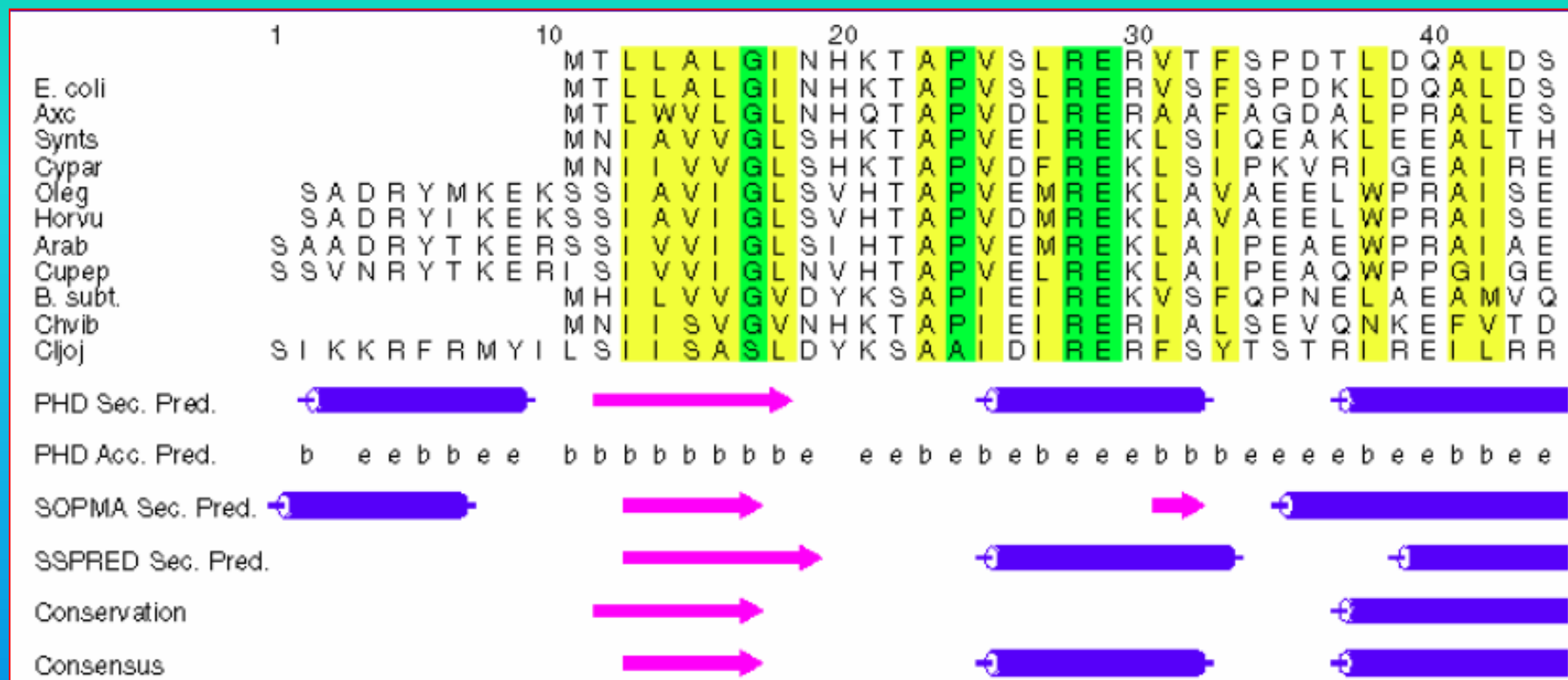
**I risultati che si ottengono sono sottoposti ad una analisi statistica per valutare l'attendibilità delle predizioni per ogni residuo.**

## PSIPRED

Il sistema di calcolo è composto da due reti neurali che analizzano i risultati della prima interazione di PSI-BLAST.

- 1** - Esecuzione di PSI-BLAST con la query desiderata
- 2** - Generazione di una PSSM (matrice posizionale di scoring) dai risultati della prima iterazione
- 3** - Predizione della struttura secondaria con una rete neurale opportunamente addestrata
- 4** - Una seconda rete neurale di correzione filtra il risultato e genera l'output definitivo, valutando la confidenza per ogni residuo.

# Consenso di metodi



**JPRED** utilizza più metodi di predizione sulla proteina query e costruisce una predizione finale mediante il confronto dei risultati dei singoli metodi.

**Esempio:**

**Proteina PDB 1FXI\_A**

**PHD:** 78.12

**DSC:** 83.33

**Predator:** 72.92

**Mulpred:**76.04

**NNSP:** 77.08

**Zpred:** 58.33

**JPRED (Consensus): 81.25**

## Calcolo dell'affidabilità delle predizioni:

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

**Un metodo più rigoroso: calcolare il coefficiente di correlazione per ogni classe di strutture secondarie:**

ad es. per le eliche

$$C_h = \frac{ab - cd}{\sqrt{(a+c)(a+d)(b+c)(b+d)}}$$

**a:** numero di residui assegnati correttamente alle eliche

**b:** numero di residui assegnati correttamente a non eliche

**c:** numero di residui assegnati in modo errato a eliche

**d:** numero di residui assegnati in modo errato a non eliche

## **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)**

## 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](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](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



# Chou and Fasman Prediction

## Hydropathy/Secondary-Structure/seg

[Search Databases with FASTA](#)

[Statistical Significance from Shuffles \(prss/prfx\)](#)

[Find Internal Duplications \(lalign/plalign\)](#)

The Kyte-Doolittle, Garnier-Osguthorpe-Robson, and Chou-Fasman programs are available for teaching purposes; much better transmembrane prediction and secondary prediction programs are available.

[memstat](#) and [TMpred](#) are more accurate transmembrane predictors

[psipred](#) and [PredictProtein](#) produce much more accurate secondary structure predictions.

## Choose:

(A) Program, (B) Protein (sequence/accession) (C) Analyze protein:

(A) Program:

(B) Protein sequence:

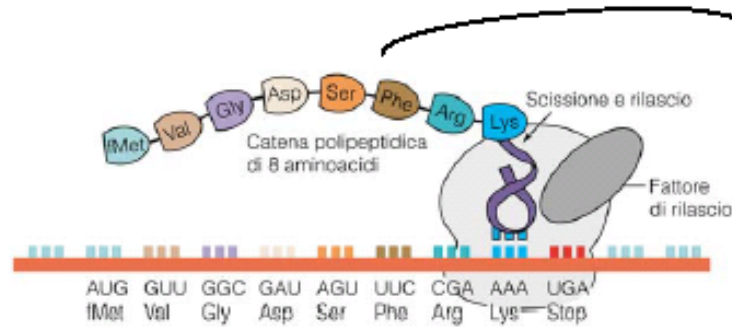
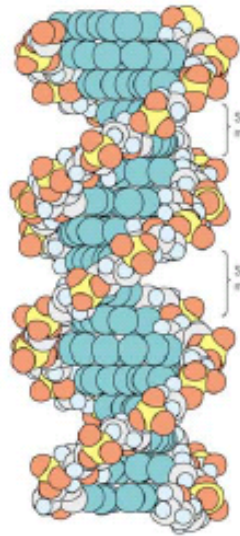
[Entrez protein sequence browser](#)

(C) Do Analysis

# Struttura terziaria

# Sequenza → Struttura → Funzione

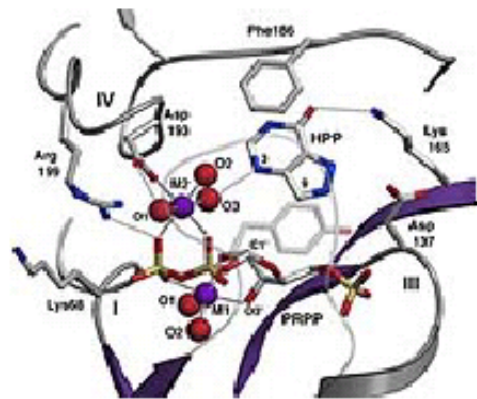
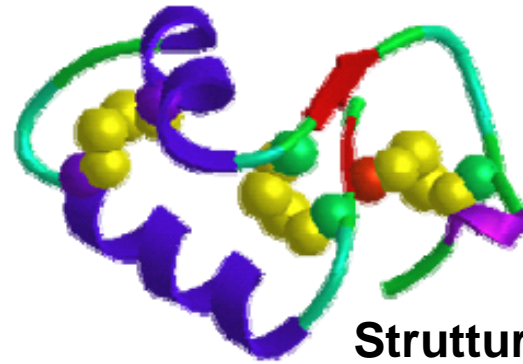
## Genomi



## Sequenza

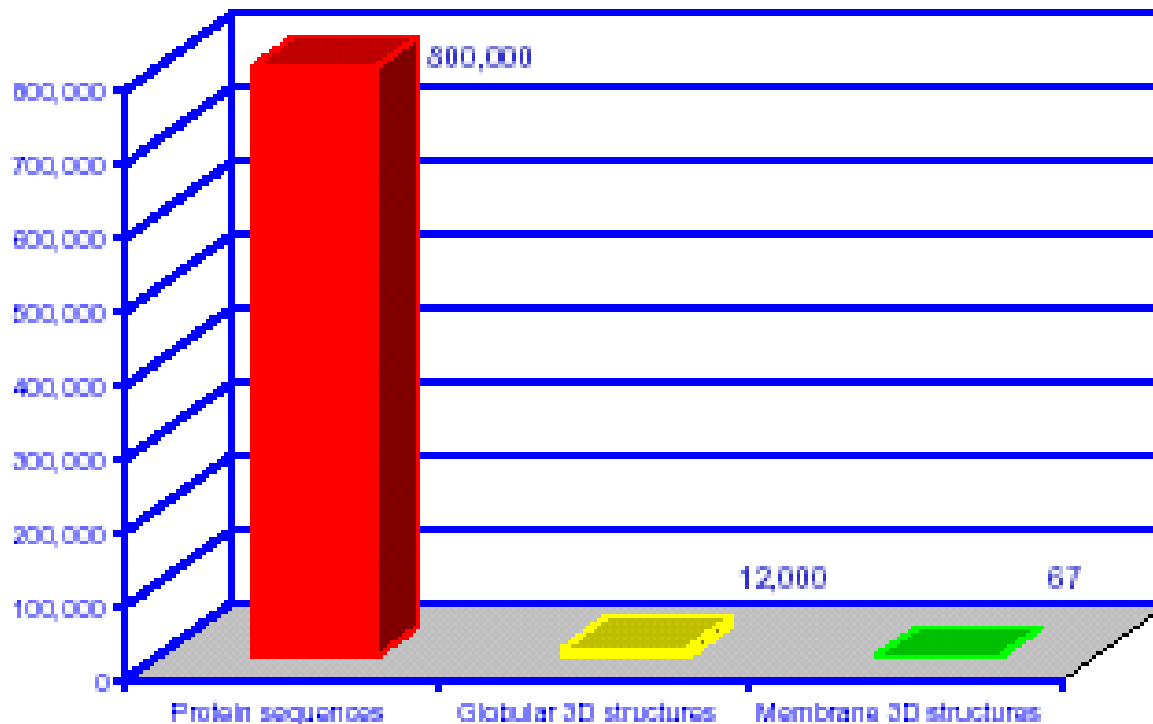


## Struttura



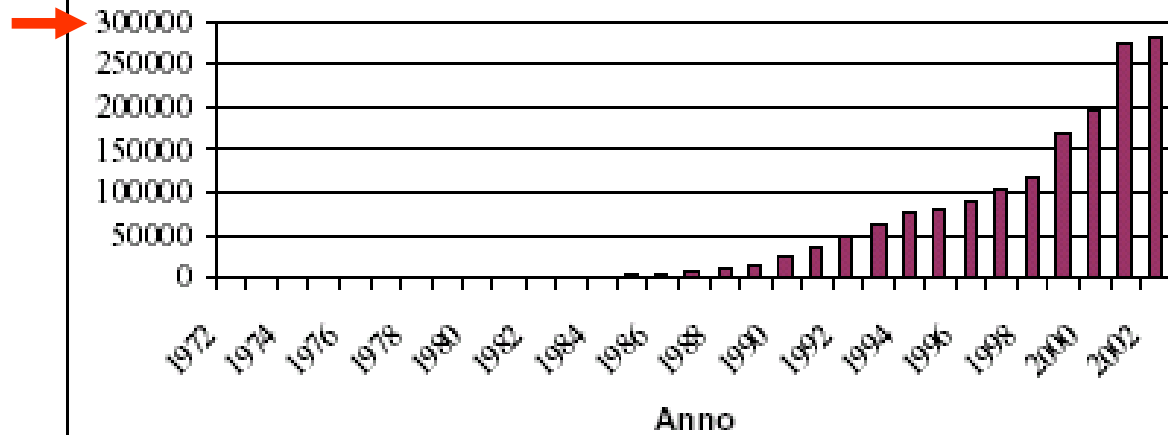
## Funzione

## Perche' predire la struttura 3D di proteine?

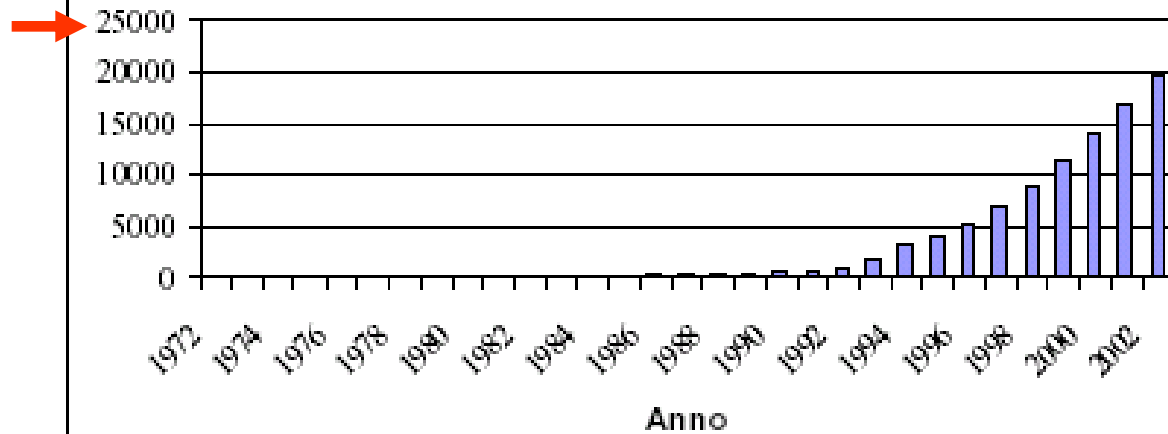


1. Studio dei meccanismi biochimici
2. Studio dei meccanismi molecolari di interazione (efficacia, ADME-assorbimento, distribuzione, metabolismo, escrezione)
3. Rational drug design

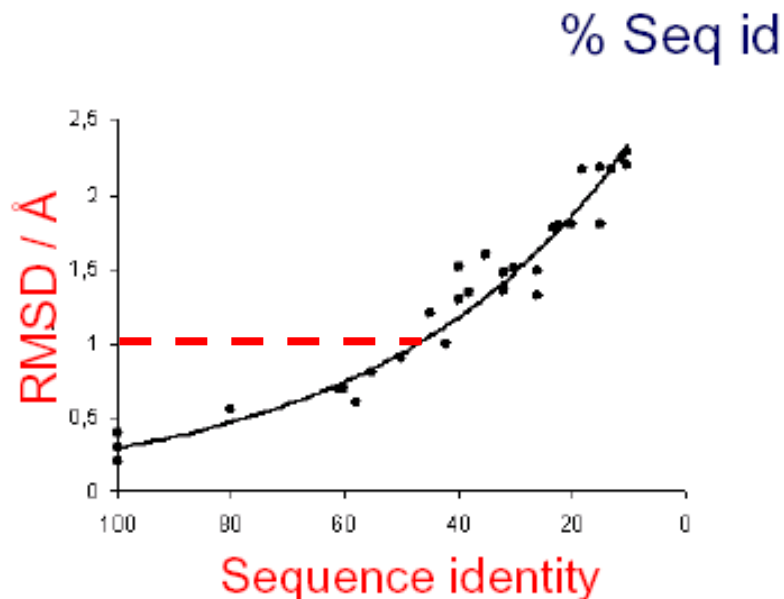
Numero di sequenze di proteine depositate nella banca dati



Numero di strutture di proteine depositate nella banca dati



Esiste una relazione (non biunivoca) fra similarita' in sequenza e similarita' in struttura



- Seq. Id. > 50%: regione "core" ~ 90% della struttura e r.m.s.d. della catena principale intorno a 1.0 Å
- Seq. Id. < 20%: regione "core" ~ 50% della struttura e r.m.s.d. della catena principale intorno a 1.8 Å

Chothia & Lesk, *EMBO Journal* (1986) vol 5, pag. 625-630

## 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.

# Struttura 3D delle proteine

## Metodi Sperimentali

- Diffrazione ai Raggi X (RX)
- Risonanza Magnetica Nucleare (NMR)

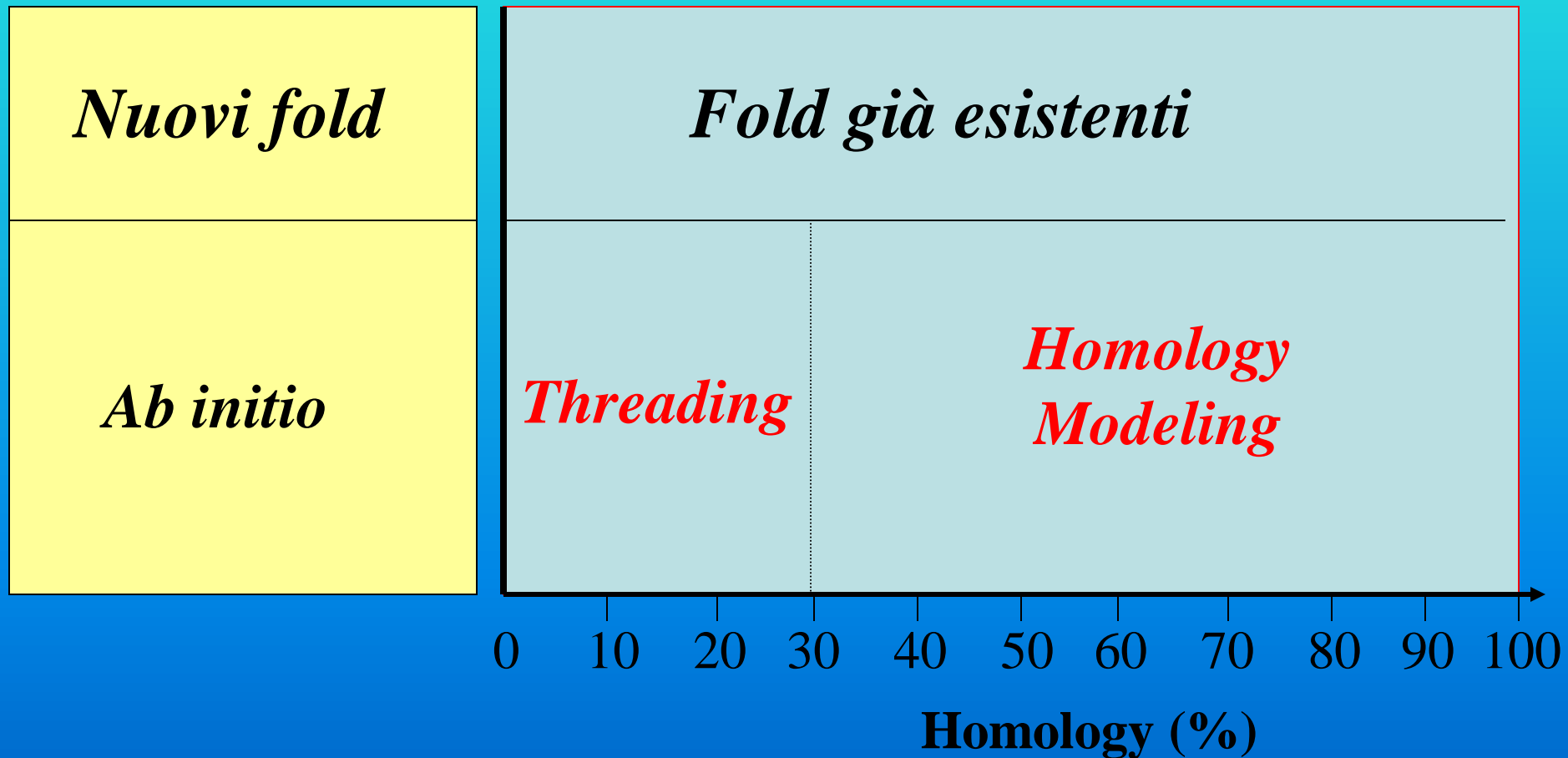
## Metodi Computazionali

- Fold Recognition
- Folding *ab-initio*

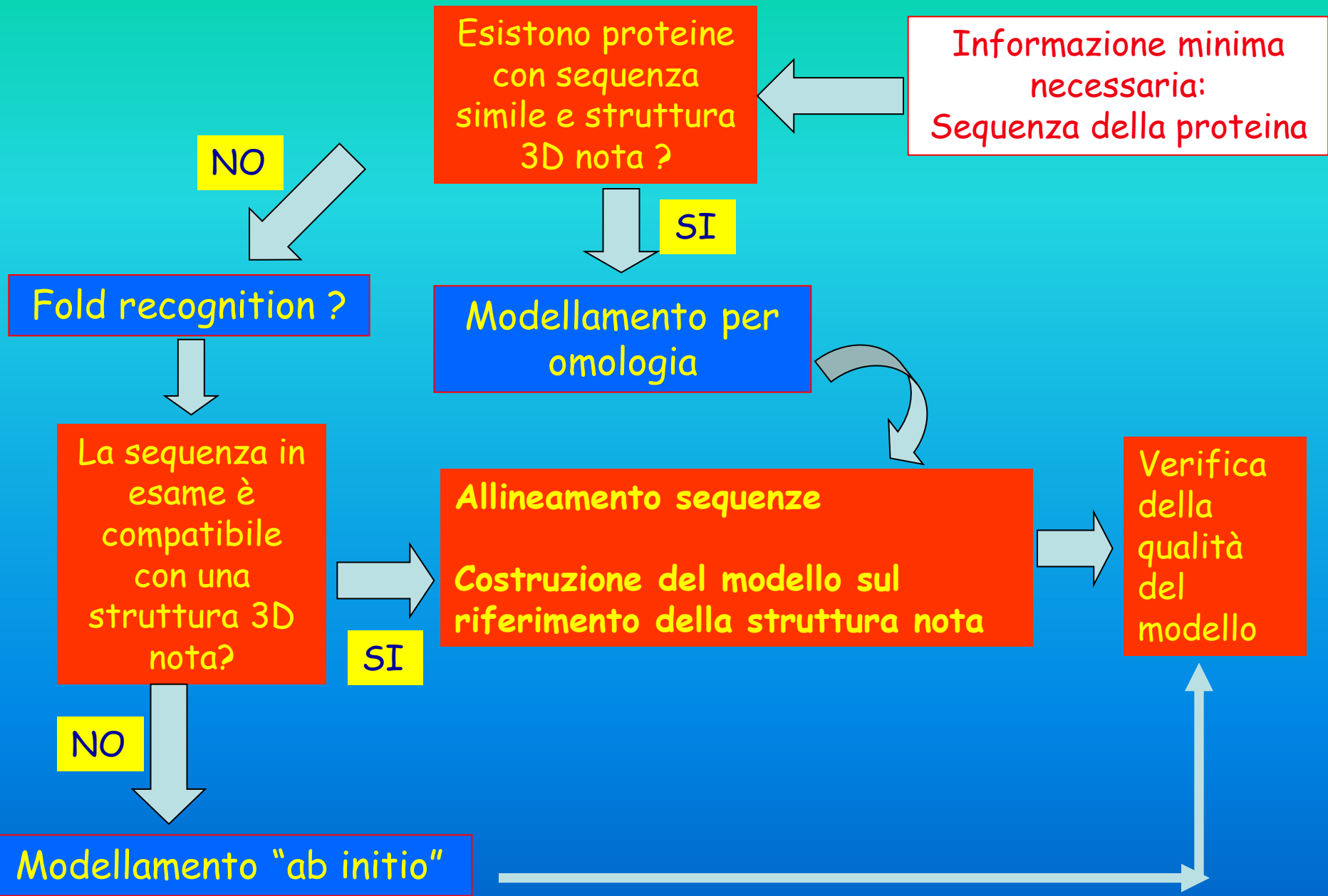
**HOMOLOGY MODELLING**



# Predizione della struttura tridimensionale di proteine



# Predizione della struttura tridimensionale delle proteine



# Modellamento per omologia

## Modellamento comparativo

- ✓ Permette di costruire il modello 3D di una proteina ('target') a partire da proteine omologhe ('template'), la cui struttura è stata caratterizzata sperimentalmente.
- ✓ La percentuale di identità di sequenza tra la proteina target e quelle template deve essere superiore al 30-40%.

**Alta identità di sequenza**



**buon allineamento delle sequenze**



**buoni modelli ottenuti per omologia**

# Modellamento Comparativo

**Modellamento delle Regioni strutturalmente conservate (SCR)**



**Modellamento delle Regioni Loop**



**Modellamento delle Catene Laterali**



**Raffinamento del modellamento**

# Modellamento Comparativo

## SEQUENZA

.....AQYSKRREVQCSVTDSEKRSLVLPNSM  
ELHAVMLQGGSDRCKVQL.....

BLAST

RICERCA DEL  
TEMPLATO

CLUSTALW

TARGET-TEMPLATE  
ALLINEAMENTO

MODELLER

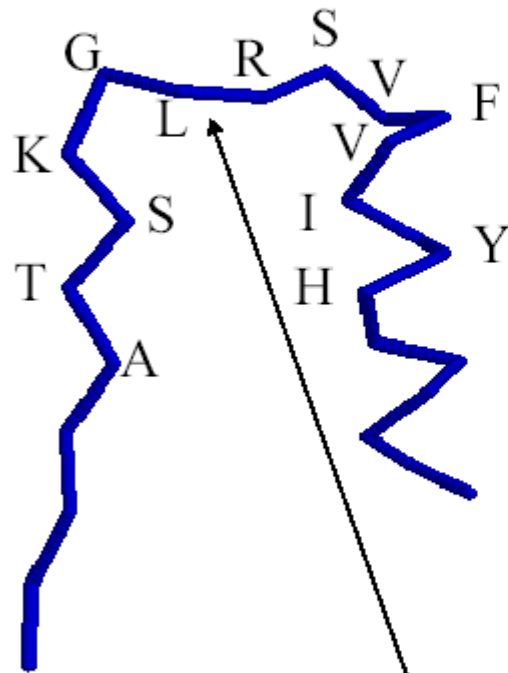
MODELLO



VALUTAZIONE DEL  
MODELLO

PROSA  
PROCHECK

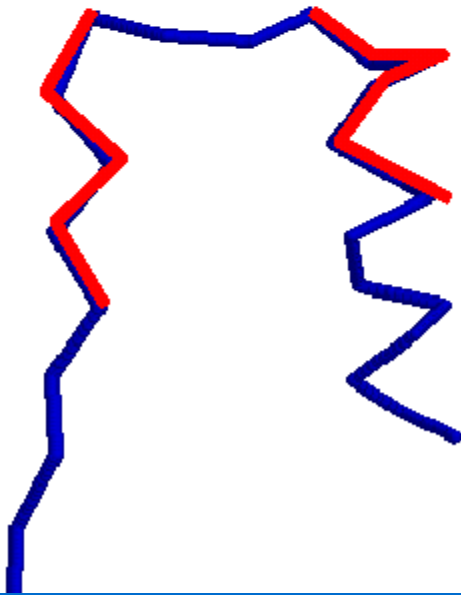
# Modeling dei loops



**ATSKGL-----RSVFVIYH**  
**ADTRGADGRATAAYVLYH**

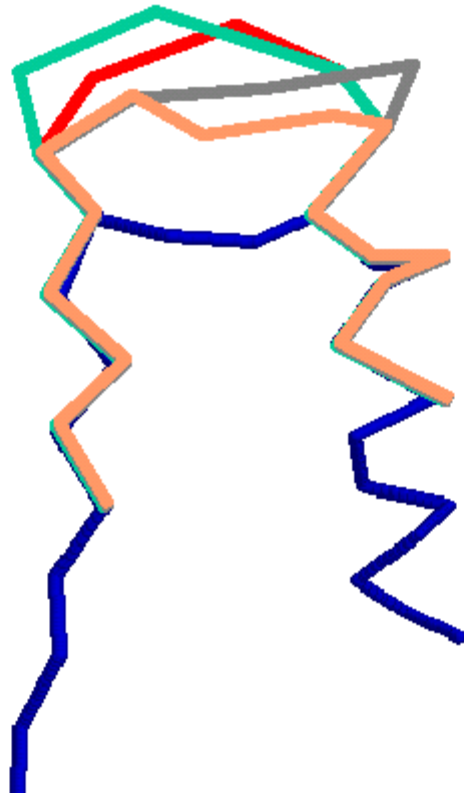
INSERIRE 4 AMINOACIDI

# Modeling dei loops





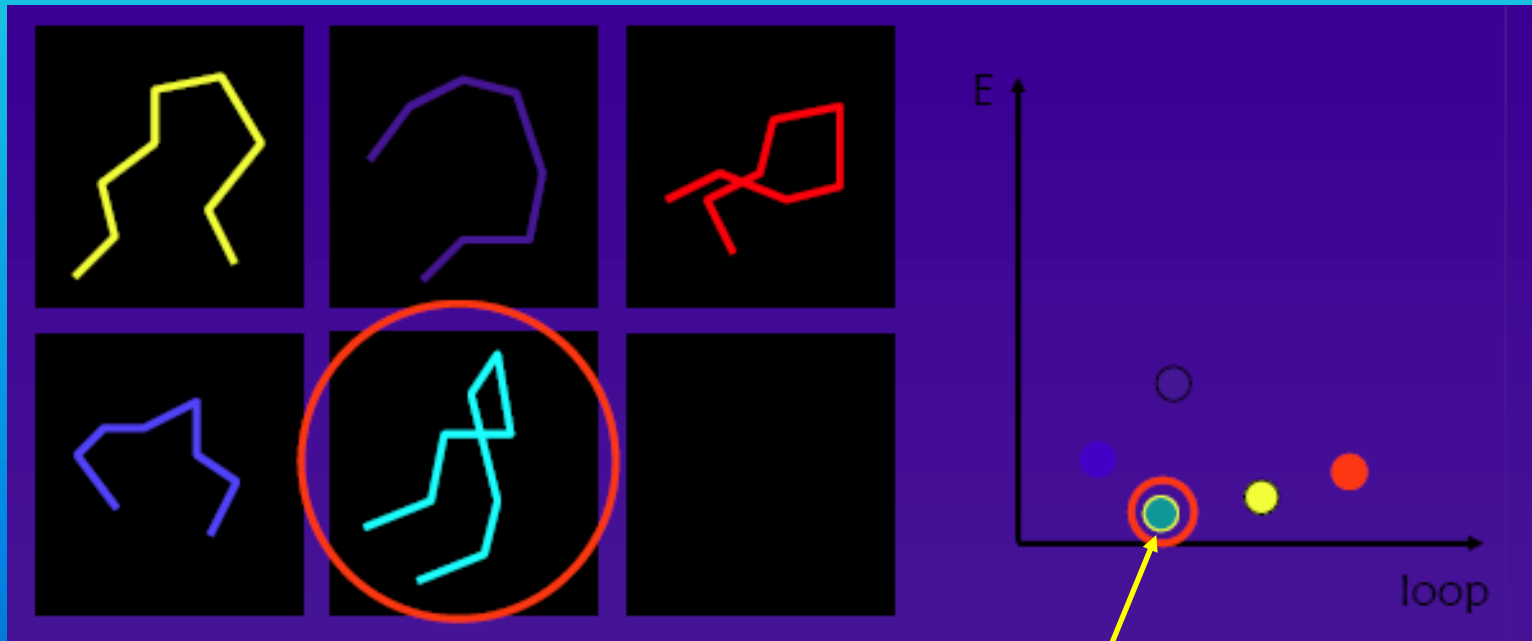
# Modeling dei loops



## Come si predice la conformazione dei loops?

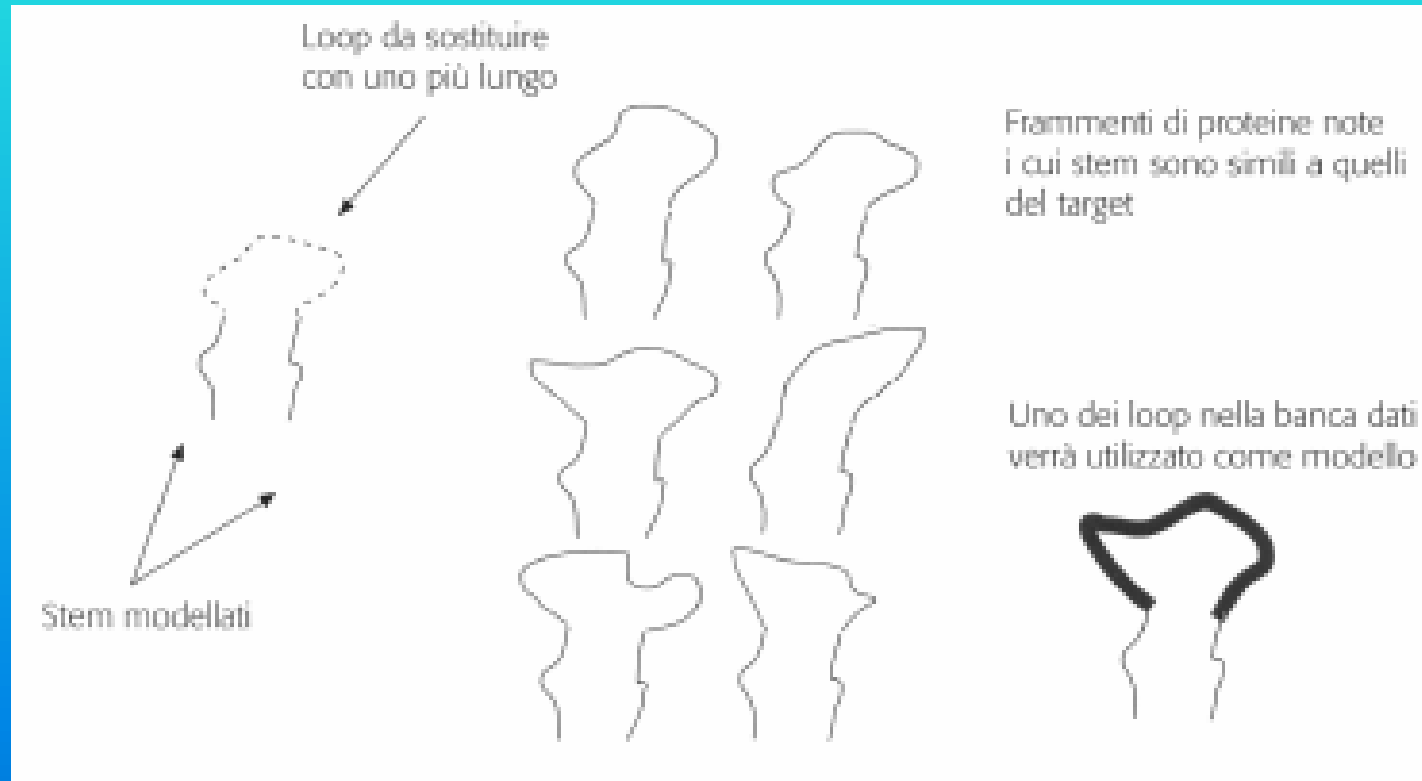
**Metodi basati sull'ottimizzazione delle conformazioni.**

Si genera un gran numero di conformazioni e si sceglie quella più adatta in termini di valori energetici.



## Metodi basati sulla ricerca nelle banche dati

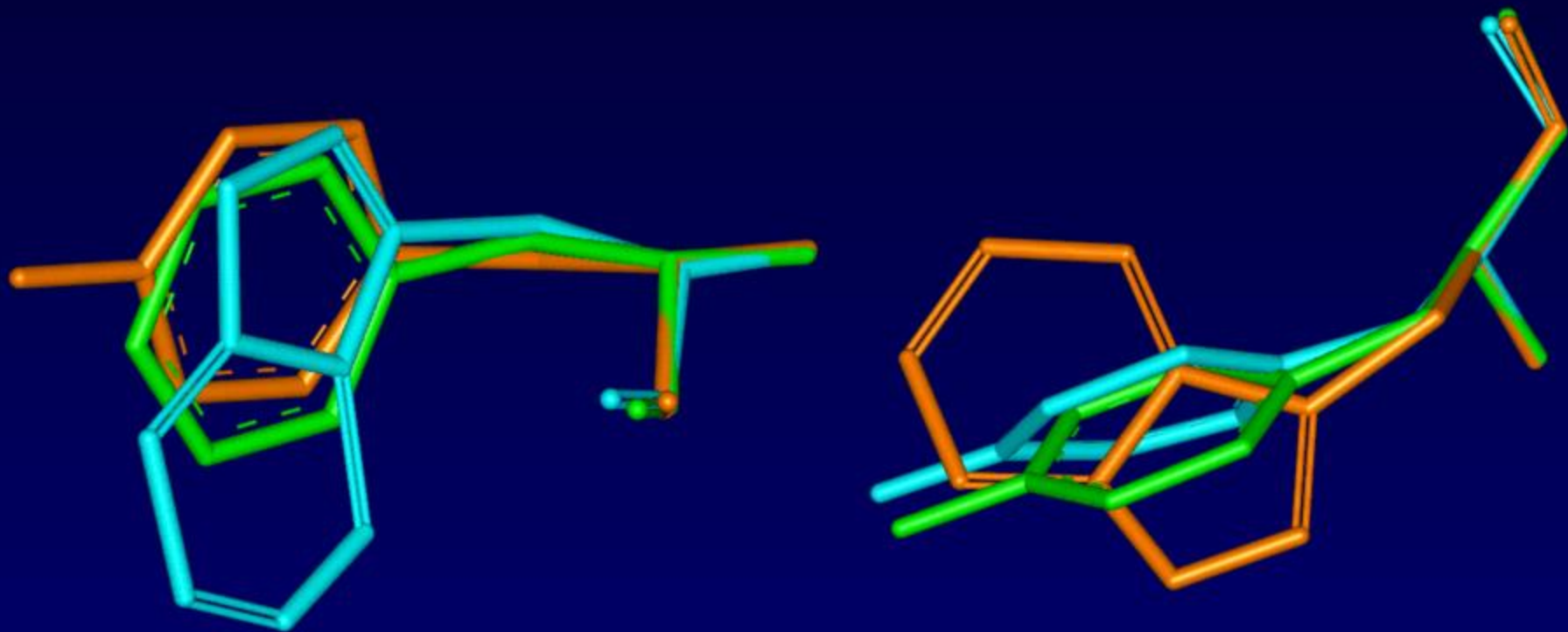
In genere all'interno di una banca dati strutturale qual è il loop che meglio si adatta alle regioni adiacenti ad esso.



Esistono anche programmi specializzati per inserire loop nelle strutture delle proteine: alcuni di essi sono liberamente fruibili e scaricabili dal sito Web degli autori.

Programs	Availability	Methods
LOOPY	<a href="http://trantor.bioe.columbia.edu/programs.html">http://trantor.bioe.columbia.edu/programs.html</a>	Colony energy with ab-initio conformation sampling and torsional space minimization
PLOP	<a href="http://francisco.compbio.ucsf.edu/~jacobson/plop_manual/plop_overview.htm">http://francisco.compbio.ucsf.edu/~jacobson/plop_manual/plop_overview.htm</a>	Extensive conformation sampling, OPLS energy, sufficient energy minimization
COILS	<a href="http://www.ch.embnet.org/software/COILS_form.html">http://www.ch.embnet.org/software/COILS_form.html</a>	Scan database of known loops from PDB.
MODELLER (loop module)	<a href="http://guitar.rockefeller.edu/modeller/modeller.html">http://guitar.rockefeller.edu/modeller/modeller.html</a>	Ab-initio conformation sampling plus CHARMM force fields
CODA	<a href="http://www-cryst.bioe.cam.ac.uk/coda/">http://www-cryst.bioe.cam.ac.uk/coda/</a>	Combine database and ab-initio approach for loop modeling

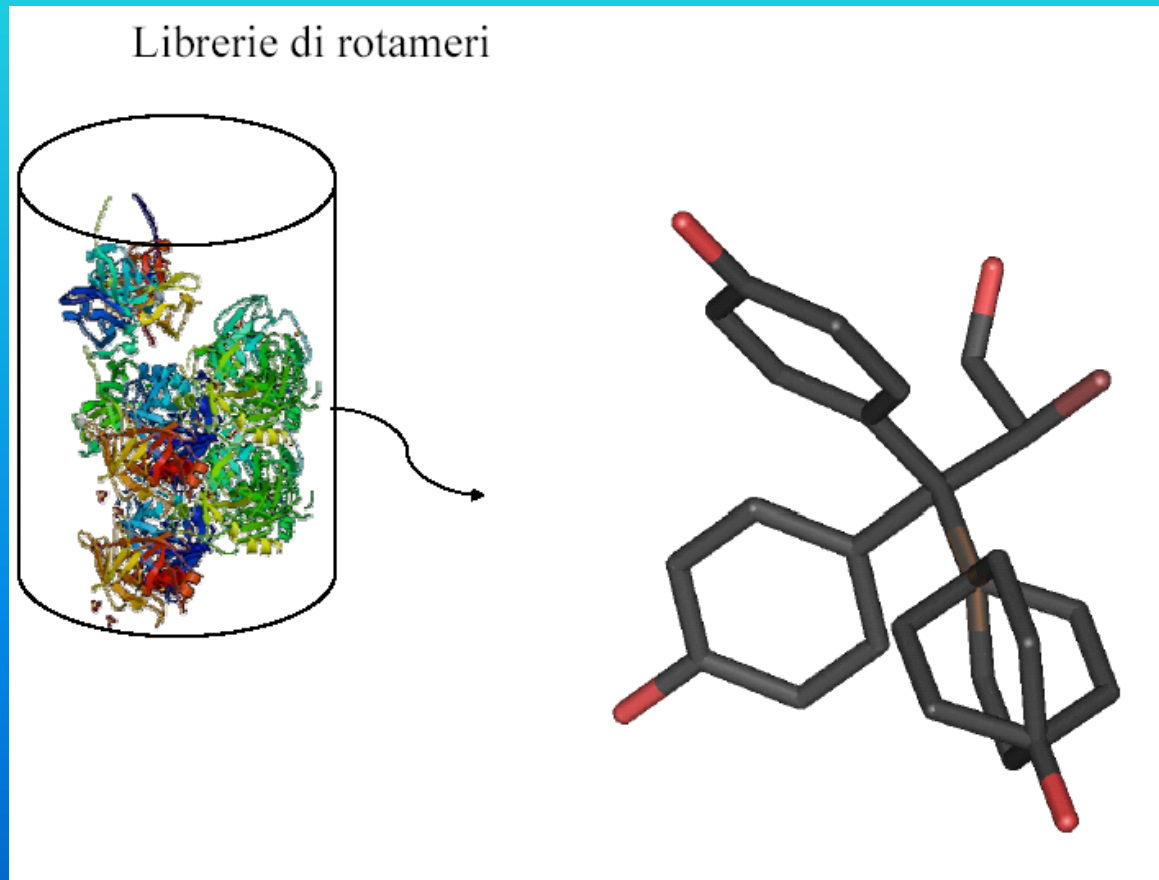
# Catene laterali



## Modellamento catene laterali

Le catene laterali degli amminoacidi hanno conformazioni energeticamente favorite, che si traducono nella frequenza con cui ogni amminoacido assume una determinata conformazione in proteine a struttura nota.

Si possono creare liste degli angoli corrispondenti alle conformazioni preferite nelle proteine note, creando le “**librerie di rotameri**”



Esistono anche programmi specializzati per inserire catene laterali nelle strutture delle proteine: alcuni di essi sono liberamente fruibili e scaricabili dal sito Web degli autori.

Programs	Availability	Methods
SCAP	<a href="http://trantor.bioc.columbia.edu/programs/jackal/">http://trantor.bioc.columbia.edu/programs/jackal/</a>	Colony energy method with simple energy and large Cartesian-coordinate rotamer library
SCWRL	<a href="http://dunbrack.fccc.edu/SCWRL3.php">http://dunbrack.fccc.edu/SCWRL3.php</a>	Simple energy with backbone-dependent rotamer library
SMOL	Contact Grishin N.V. at <a href="mailto:Nikolai.Grichine@UTSouthwestern.Edu">Nikolai.Grichine@UTSouthwestern.Edu</a>	Optimized scoring function with extended backbone-dependent rotamer library and Monte Carlo search method
SCCOMP	<a href="http://atlantis.weizmann.ac.il/~eyale/">http://atlantis.weizmann.ac.il/~eyale/</a>	Optimized scoring function and Gibbs sampling like algorithm
RAMP	<a href="http://www.ram.org/computing/ramp/ramp.html">http://www.ram.org/computing/ramp/ramp.html</a>	knowledge based potentials and small rotamer library
SMD	<a href="http://condor.urbb.jussieu.fr/Smd.php">http://condor.urbb.jussieu.fr/Smd.php</a>	Flex force field, small rotamer library and dynamic cluster analysis of known structures
Conformat	Contact Koehl P at <a href="mailto:koehl@esb.stanford.edu">koehl@esb.stanford.edu</a>	self-consistent mean field and small rotamer library
Maxsprout	<a href="http://www.ebi.ac.uk/maxsprout/">http://www.ebi.ac.uk/maxsprout/</a>	Rough energy function and small rotamer library

CASP7 Home page - Netscape Browser

File Edit View Go Bookmarks Tools Help

Swiss-model SEARCH http://predictioncenter.org/casp7/Casp7.html SECURITY CENTER


Personal Weather Wetmail Inside Netscape News Sites (40)

CASP7 Home page

*7<sup>th</sup> Community Wide Experiment on the*

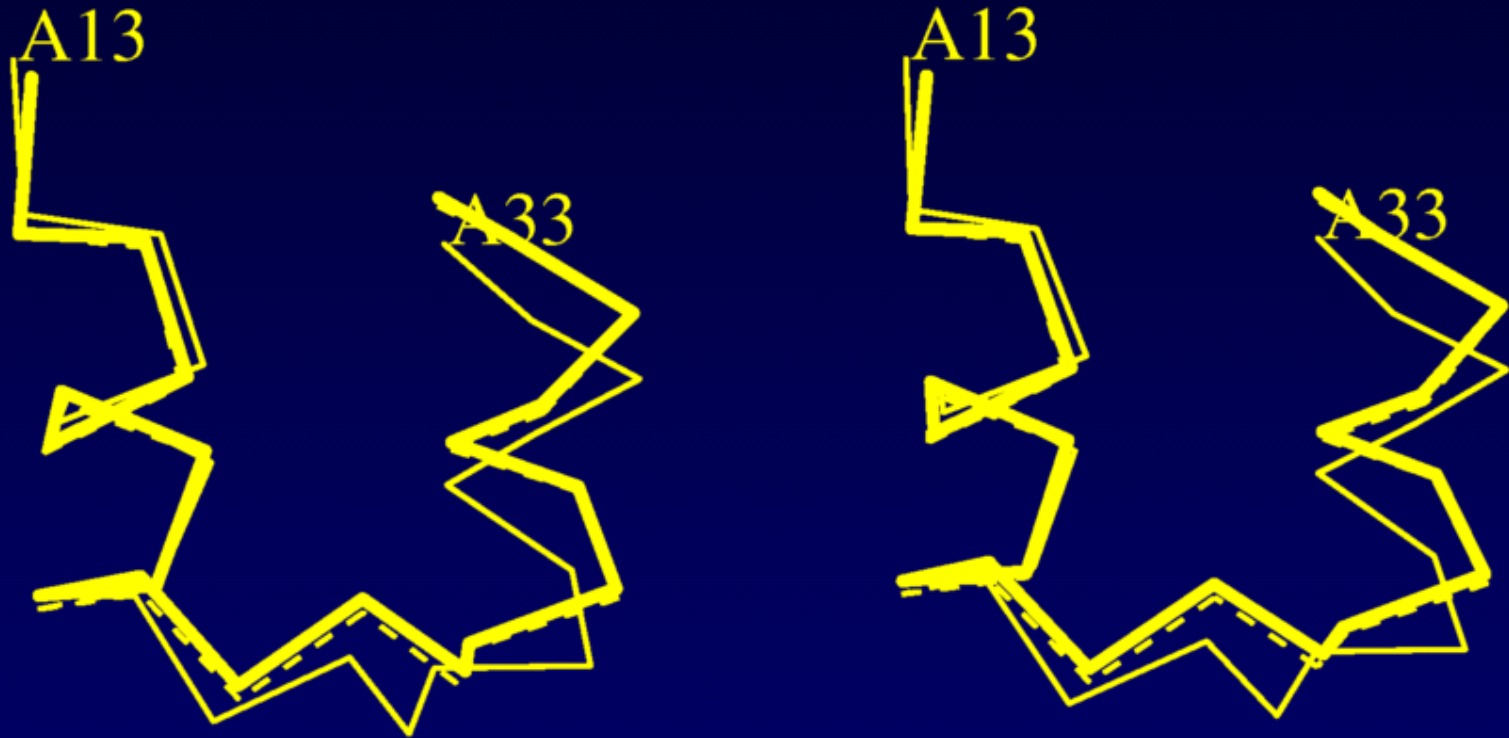
**Critical Assessment of Techniques for Protein Structure Prediction**

*Asilomar Conference Center, Pacific Grove, CA  
November 26-30, 2006*

The logo for CASP7 is enclosed in a black rectangular border. On the left side, the letters 'C', 'A', 'S', 'P', and '7' are stacked vertically in a bold, black, sans-serif font. To the right of the text is a stylized illustration of a protein structure, showing a red ribbon-like backbone and a white, semi-transparent surface representing the protein's fold.



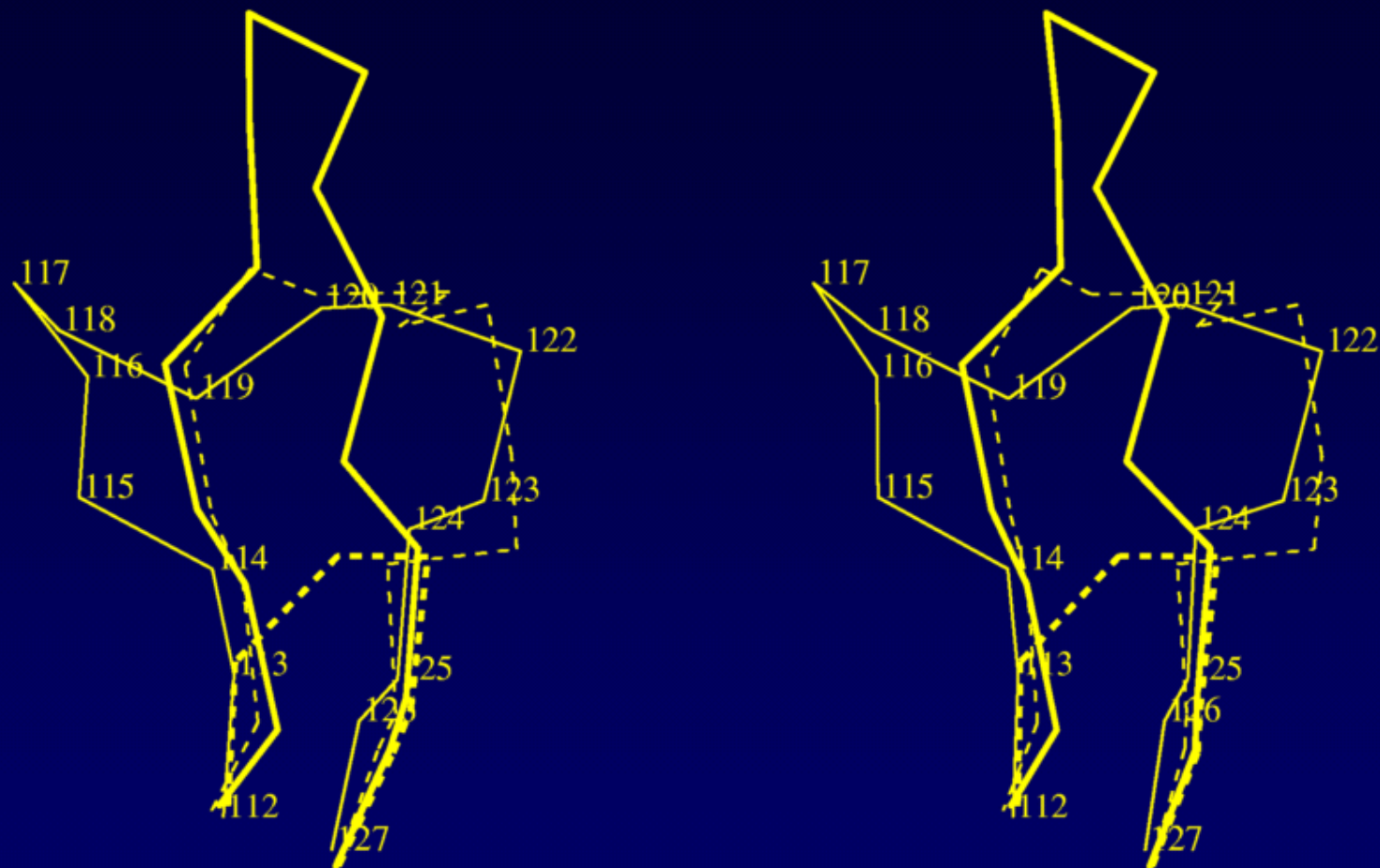
# Modeller: errori nella catena polipeptidica



Sono mostrate le distorsioni in regioni correttamente allineate.

Linea sottile: struttura ai Raggi X; Linea spessa: template utilizzato per il modellamento; Linea tratteggiata: modello ottenuto per omologia.

# Modeller: errori nelle regioni non allineate



**Errori nelle regioni allineate male.**

**Linea sottile: struttura ai Raggi X; Linea spessa: modello ottenuto per omologia; Linea tratteggiata: template utilizzato per il modellamento.**

# Modeller: errori nell'allineamento di sequenze

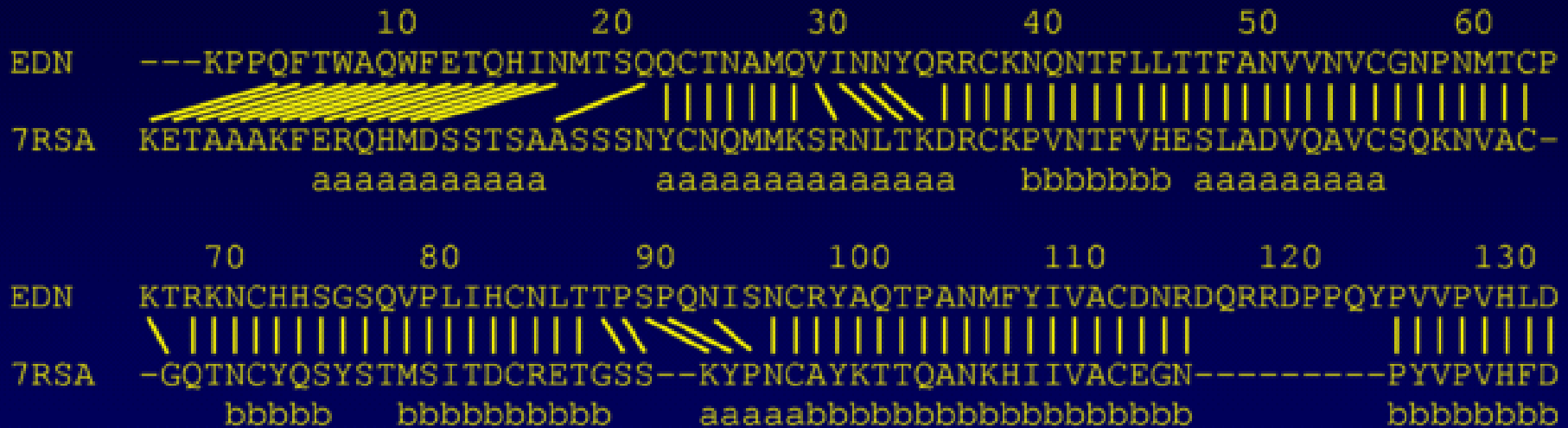
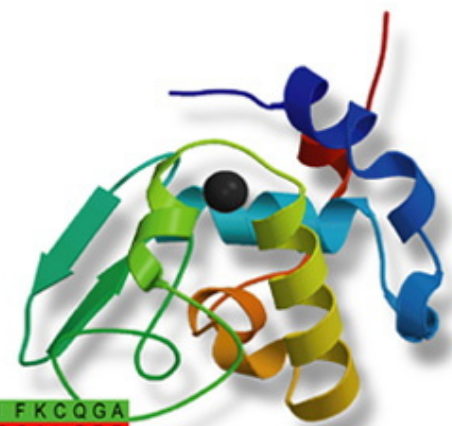


Figure 7: Errors in the sequence alignment of human eosinophil neurotoxin and ribonuclease A. Automatically derived sequence alignment is shown. The black lines show correct equivalences, that is residues whose  $C_{\alpha}$  atoms are within  $5\text{\AA}$  of each other in the optimal least-squares superposition of the two X-ray structures. The bottom line indicates helices (a) and strands (b), as assigned in the human eosinophil neurotoxin structure by program DSSP [83]. Reprinted with permission from [38].

# Modeller

Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints



```

A I L V G S M P R R D G M E R K D L L K A N V K I F K C Q G A
V E V C P V D C F Y E G P N F L V I H P D E C I D C A L C E P
G A C K P E C P V N I I Q G S - - Y A I D A D S C I D C G S
C - - I A C G A C K P E C P V N I I Q G S - - I Y A I D A D S
    
```

About MODELLER
News
Download & Installation
Registration
Discussion Forum
User Manual

## About MODELLER

MODELLER is used for homology or comparative modeling of protein three-dimensional structures (1). The user provides an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms. MODELLER implements comparative protein structure modeling by satisfaction of spatial restraints (2, 3), and can perform many additional tasks, including de novo modeling of oligopeptides, optimization of various models of protein structure with respect to a flexibly defined objective function, multiple alignment of protein sequences and/or structures, clustering, searching of sequence databases, comparison of protein structures, etc. MODELLER is written in Fortran-90 and is meant to run on a UNIX system.

1. M.A. Marti-Renom, A. Stuart, A. Fiser, R. Sánchez, F. Melo, A. Sali. Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. 29, 291-325, 2000.

2. A. Sali & T.L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815, 1993

Please send MODELLER bug reports and suggestions to [bozidar@quitar.rockefeller.edu](mailto:bozidar@quitar.rockefeller.edu)



## Database of Comparative Protein Structure Models

Welcome to MODBASE, a database of three-dimensional protein models calculated by [comparative modeling](#).

### About MODBASE

[General Information](#)

[Glossary](#)

[Authors and acknowledgements](#)

[Publications](#)

[Related resources](#)

Users of MODBASE are requested to cite this article in their publications:  
[MODBASE, a database of annotated comparative protein structure models.](#)  
 Ursula Pieper, Narayanan Eswar, Ashley C. Stuart, Valentin A. Ilyin, Andrej Sali.  
*Nucl. Acids Res.* **30**, 255-259, 2002.

### MODBASE Contents

837,698 [Reliable Models](#) or [PSI-BLAST Fold Assignments](#) for domains in 415,937 proteins. Last Update on 04/03/02. MODBASE [statistics](#).

### Search for Models

Enter SwissProt/TrEMBL/GenBank/PDB identifier or descriptor:



### [ADVANCED SEARCH](#)

Login

[HELP](#)

[ACADEMIC LOGIN](#)

[USER LOGIN](#)

[LOGOUT](#)

Current logins: *public*.

Some datasets are accessible freely without a login (ie, the "public" model set). Some datasets are available to academic users only (ie, our "SP/TR" model set). And some datasets require a specific username and password. For commercial access to the models.