

PREDICCIÓ *AB INITIO* DE PLEGAMENTS

ESTHER PEÑA

ALÍCIA MOYA GARRIDO

GUILLEM BRUGUERA

ÍNDEX DE CONTINGUTS

1. Introducció

1. Mètodes de predicció d'estructura terciària
2. Predicció *ab initio*

2. Situació actual → principals problemes

3. Possibles aproximacions:

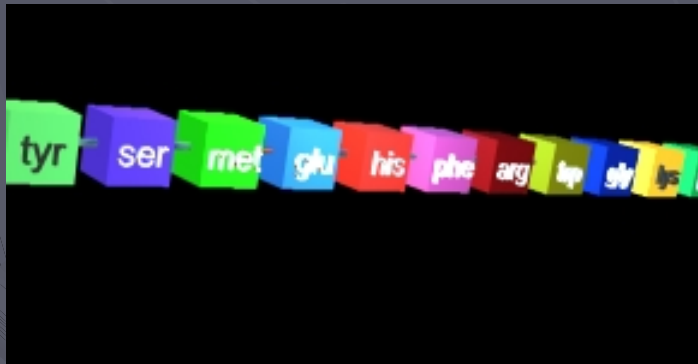
1. Models de complexitat reduïda
2. Potencials

ÍNDIX DE CONTINGUTS

- 5. Descripció dels principals mètodes:
 - 1. Folding Thermodynamics of Peptides
 - 2. Monte Carlo – Metropolis
 - 3. Contact maps
 - 4. Massive parallel computing
 - 5. Rosetta
- 6. CASP
- 7. Aplicacions
- 8. Bibliografia

INTRODUCCIÓ

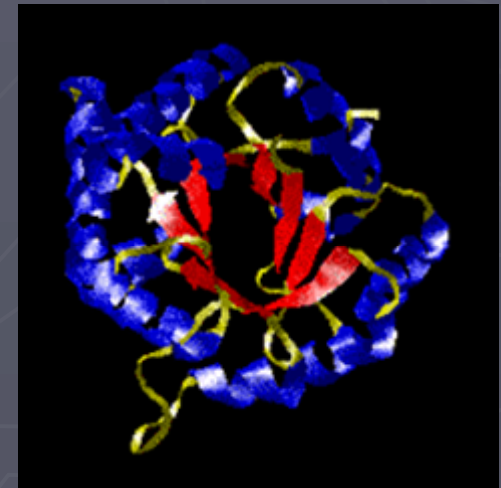
? Mètodes de predicció de l'estructura terciària:



Comparative modelling

Fold recognition (threading)

Ab initio prediction



INTRODUCCIÓ

? Predicció *ab initio* :

- Objectiu: Donada la seqüència d'aminoàcids d'una proteïna, deduir l'estructura del seu estat nadiu.

- Principi → **HIPÒTESI TERMODINÀMICA**

“l'estructura amb que una proteïna es plega és aquella amb l'energia lliure global més baixa”

INTRODUCCIÓ

- Requisites:

- ? Trobar la funció que permeti calcular l'energia lliure d'una estructura → que tingui en compte totes les energies que participen en el seu plegament

Funció que ens permeti distingir les estructures correctes de les incorrectes

- ? Desenvolupar un protocol eficient per trobar l'estructura amb la mínima energia lliure
- ? Bon mètode de cerca per explorar l'espai conformacional

SITUACIÓ ACTUAL: principals problemes

- ? És difícil trobar la funció “perfecte”
- ? Es necessita una enorme quantitat de recursos computacionals per resoldre una funció d'aquest tipus

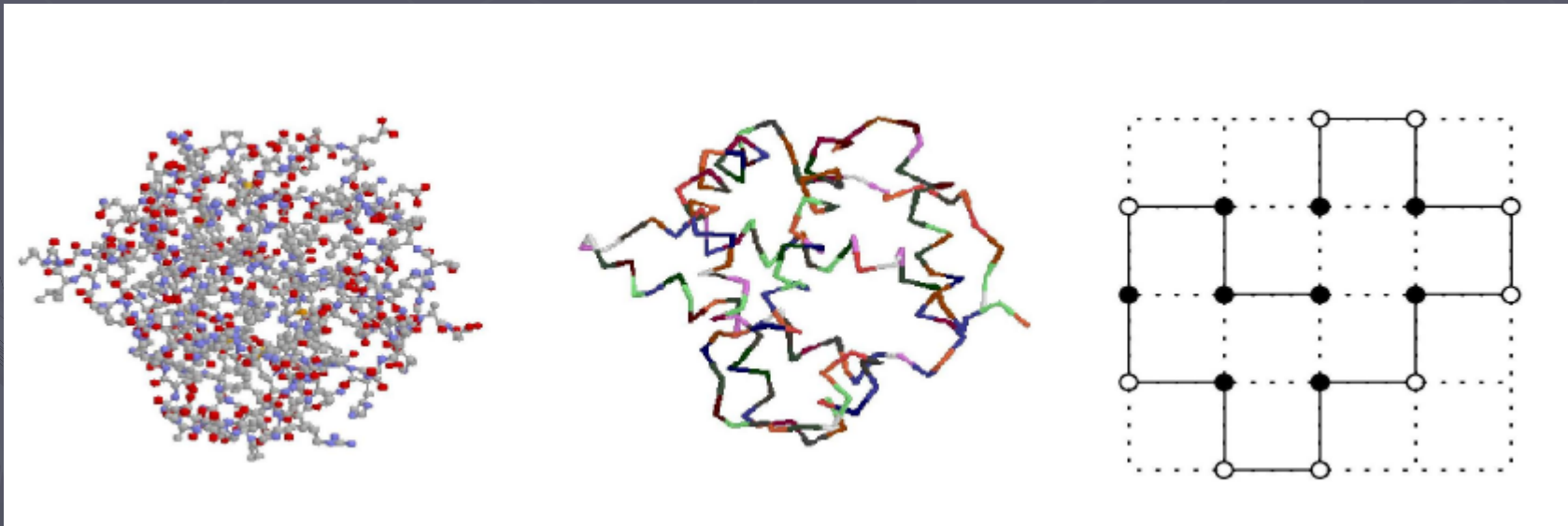
SITUACIÓ ACTUAL: principals problemes

Cap dels mètodes actuals és capaç de
predir acuradament tots els tipus de
proteïnes

- La majoria funcionen pitjor amb “all-beta proteins”
- Tots els mètodes fallen quan s'enfronten a seqüències de més de 150 residus

POSSIBLES APROXIMACIONS

? Models de complexitat reduïda → lattice models :
Hp model



POSSIBLES APROXIMACIONS

? Models de complexitat reduïda → lattice models : Hp model

■ Els aminoàcids es poden dividir en :

? Hidrofòbics (H)

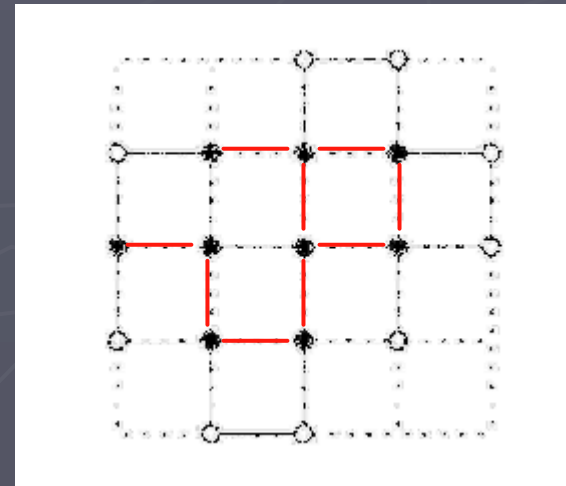
? Polars (P)

● hydrophobic amino acid

○ hydrophilic amino acid

— Covalent bond

— H-H contact



L'objectiu es maximitzar el nombre d'enllaços H-H

POSSIBLES APROXIMACIONS

- ? Models de complexitat reduïda → lattice models : Hp model
- ? L'energia lliure de la conformació es defineix com el número negatiu de contactes hidrofòbic-hidrofòbic consecutius .
 - Un contacte es defineix com dos monòmers no consecutius en la cadena, ocupant posicions adjacents a l'enreixat
 - **Problema**: la optimització es basa en la posició dels residus hidrofòbics i per tant els polars tenen massa llibertat i poden adoptar disposicions poc naturals si aquests segments són massa llargs o es troben als extrems.

POSSIBLES APROXIMACIONS

- ? Es redueix el nombre de càlculs en limitar els graus de llibertat
- ? Limitem els àtoms dels C-aa posicions en un enreixat 3D o 2D
- ? La seqüència de la proteïna es representa com a un camí a través d'un enreixat
- ? S'emfatitza la formació d'un nucli hidrofòbic

ESTRATÈGIES

? Models de complexitat reduïda → lattice models

- Tot i ser computacional i analíticament molt simples , la seva baixa resolució (5-20 Å) és un problema important
- El seu principal inconvenient és la presència d'errors sistemàtics relacionats amb l'estructura secundària:
 - ? Biaixos
 - ? Dificultat de representar correctament hèlix

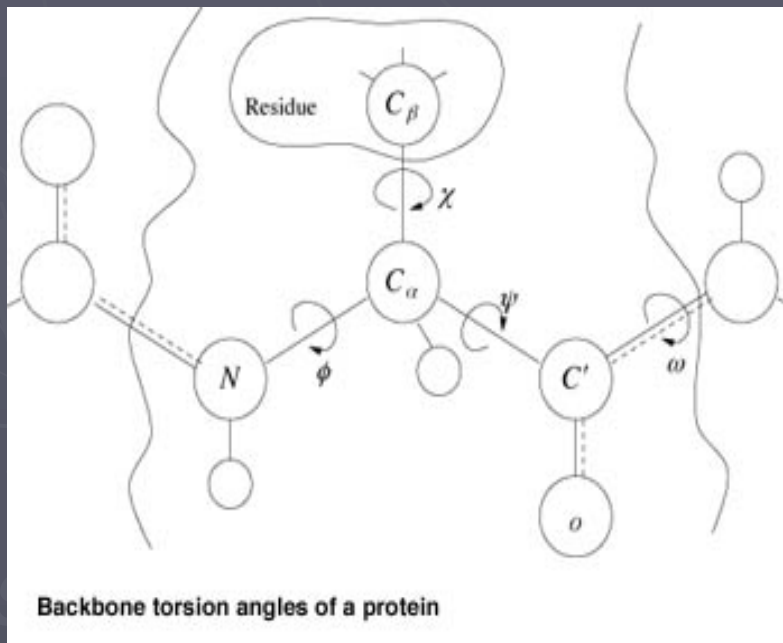
POSSIBLES APROXIMACIONS

? Models de complexitat reduïda → discrete off-lattice models

- Derivem un “lattice” (enreixat) on al centre es troba la nostra proteïna.
- Aquest lattice constitueix un “espai” limitat on la nostra proteïna es pot plegar
- Per construir aquest “espai” limitem el nombre de valors que els angles de torsió de l'esquelet principal de la nostra proteïna poden prendre per a cada residu .

ESTRATÈGIES

? Models de complexitat reduïda → discrete off-lattice models



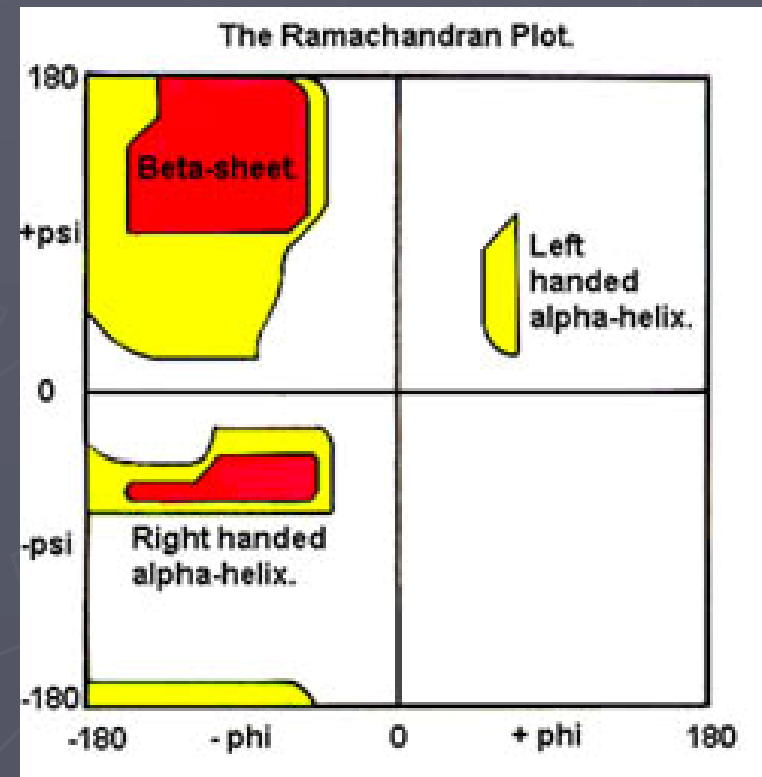
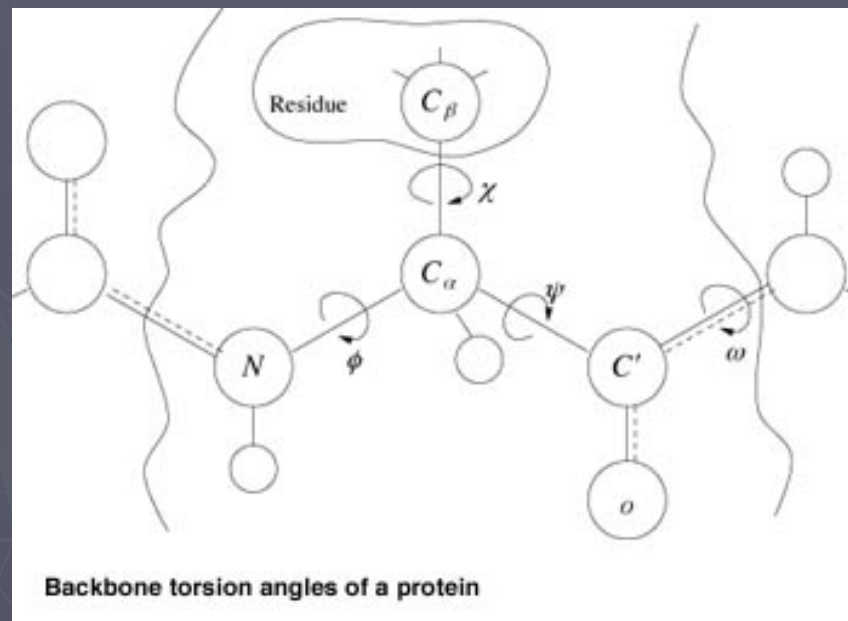
L'enllaç peptídic és rígid

L'enllaç CO-NH té caràcter de doble enllaç

L'angle de torsió omega té generalment només dues conformacions: cis (0°) i **trans** (180°)

ESTRATÈGIES

- ? Models de complexitat reduïda → discrete off-lattice models



Els altres dos angles de torsió (phi i psi) també estan altament restringits

ESTRATÈGIES

- ? Models de complexitat reduïda → discrete off- lattice models
 - Els valors dels angles de torsió són discrets:
 - ? $\{0,180\}$ per l'angle omega
 - ? Segons el gràfic de Ramachandran pels angles psi i phi



Espai amb $(2k)^n$ possibles conformacions

N = nombre d'aminoàcids

K = nombre de valors diferents que poden adoptar els angles psi i phi

ESTRATÈGIES

? Avaluació de l'energia lliure:

- Potencials basats en mecànica molecular (físics)
- Potencials estadístics (basats en l'estructura de proteïnes de la PDB) → alineaments de seqüència múltiples

POTENCIALS FÍSICS

- “Estudi de les interaccions interatòmiques”:
 - Descobriment de l'estructura tridimensional de les proteïnes basat únicament en:
 - Longitud d'enllaços
 - Flexió d'angles
 - Interaccions torsionals
 - Interaccions no covalents

POTENCIALS ESTADÍSTICS

Kolinski, A. (2004) Protein modeling and structure prediction with a reduced representation

? Restriccions de curt abast :

- Independents de seqüència
- Dependents de seqüència

? Restriccions d'ampli abast:

- Independents de seqüència
- Dependents de seqüència

POTENCIALS ESTADÍSTICS

? Restriccions genèriques de curt abast
independents de la seqüència:

1. Restriccions de l'angle planar del carboni alfa
2. Biaix cap a la rigidesa d'un fragment de la cadena
3. Biaixos cap a estructures secundàries regulars
4. Biaixos en contra d' estructures massa plegades (crumpled)

POTENCIALS ESTADÍSTICS

? Com apliquem els potencials estadístics a la predicció *ab initio* ?

- Cadascuna de les restriccions es representen mitjançant unes fórmules que són “**factors de correcció**”
- L'estructura d'aquestes fórmules és:

? $B(?) = \text{número (0,5 - 4.0)} \times f \times e_g$

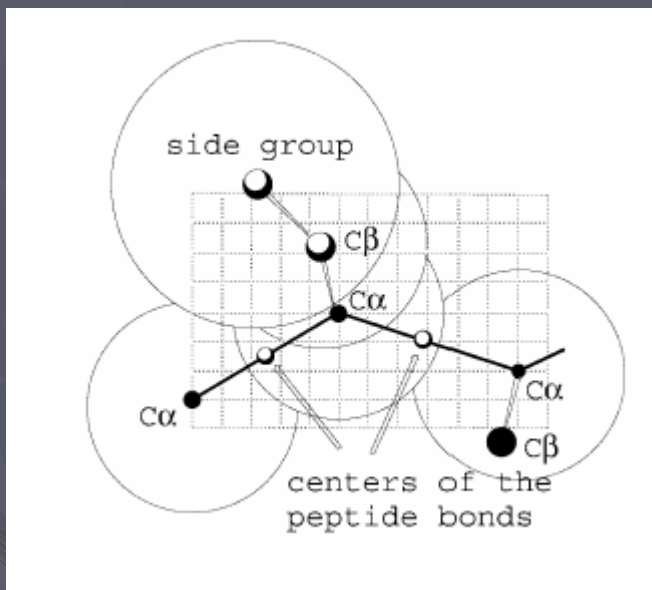
Tipus de biaix

Defineix la propensitat a certa estructura

Factor escalant comú a tots els biaixos de curt abast genèrics

POTENCIALS ESTADÍSTICS

? Restriccions genèriques de curt abast independents de la seqüència



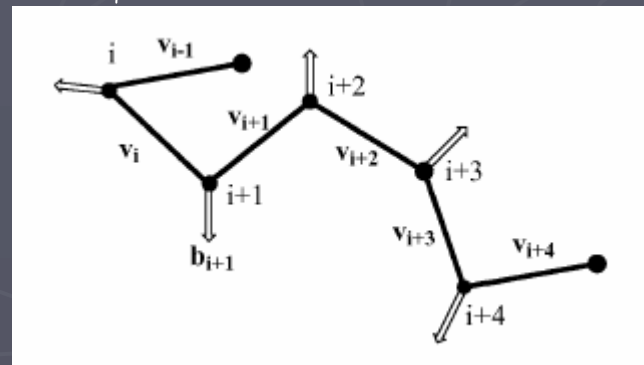
Model CABS

1. Restriccions de l'angle planar del carboni alfa

$$4.1 \text{ \AA} \leq |r_{i+2} - r_i| \leq 7.4 \text{ \AA}$$

Rang dels angles planars:

70° -150°



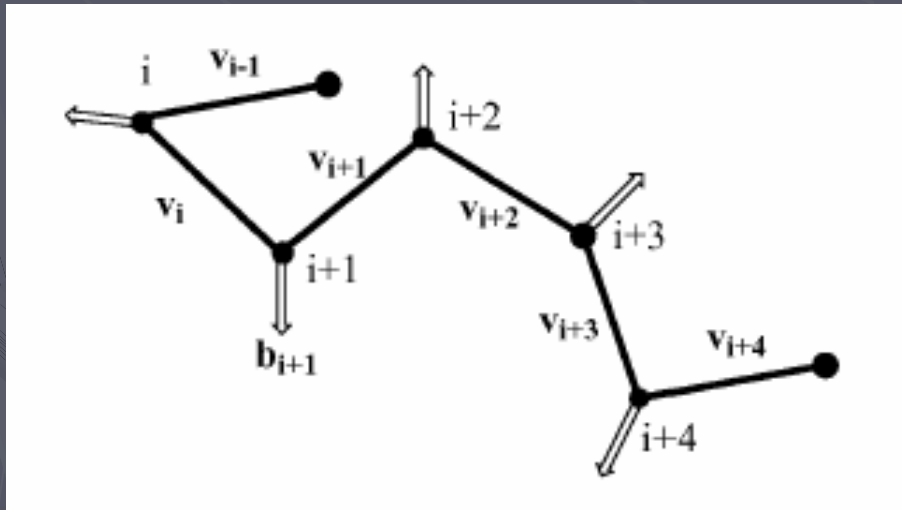
On : i-th → Cα
 v_i → enllaç corresponent
 r_i → coordenades cartesianes

POTENCIALS ESTADÍSTICS

? Restriccions genèriques de curt abast independents de la seqüència

2. Biaix cap a la rigidesa d'un fragment de la cadena

La distribució de les distàncies “end-to-end” és de tipus Gaussià



En les proteïnes això es correspon amb una distribució bimodal



POTENCIALS ESTADÍSTICS

? Restriccions genèriques de curt abast independents de la seqüència

3. Biaix cap a estructures secundàries regulars

- Les hèlix són, habitualment, orientades a la dreta, amb un valor característic de distància entre els residus
- Les làmines beta tenen normalment una geometria “up –and-down”

Nota: l'aplicació de la fórmula del biaix corresponent es basa en el càlcul de les distàncies igual que en l'exemple anterior. Si utilitzem les diverses eines de les que disposem per predir l'estructura secundària de la proteïna, podem aplicar aquest biaix a zones més amples de la mateixa.

POTENCIALS ESTADÍSTICS

? Restriccions genèriques de curt abast independents de la seqüència

4. Biaixos en contra d' estructures massa plegades (crumpled)

- Les estructures on els loops que canvien la propagació de la cadena es troben molt junts són extremadament estranys a la natura.
- S'assumeix que la longitud mínima entre dos loops ha de ser com a mínim de 5 residus

POTENCIALS ESTADÍSTICS

- ? Restriccions genèriques de curt abast independents de la seqüència

$$Eg = S (BB + BS + BH + BHH + BE + BEE + BC)$$

POTENCIALS ESTADÍSTICS

? Restriccions genèriques de curt abast dependents de la seqüència:

- Freqüència relativa amb que s'observa certa geometria local donat un parell d'aminoàcids
- L'energia conformacional total s'obté a partir del càlcul independent de tres potencials diferents tenint en compte 3, 4 i 5 aa respectivament.

POTENCIALS ESTADÍSTICS

? Interaccions d'ampli abasts → repulsió dels Ca i C β

■ Interaccions de repulsió genèriques :

? Ca → C β

? Ca → centre els enllaços peptídics

? Ca → grups laterals

Table 1. Cut-off distances for the hard-core repulsive interactions

Interaction	Cut-off (Å)
C α - C α	3.05
C α - C β	3.66
C β - C β	3.05

■ Interaccions depenents de seqüència :

? Interacció del grups laterals (límit = 4,5 amstongs)

? Importància del context de la interacció

POTENCIALS ESTADÍSTICS

- Interaccions depenents de seqüència :

- ? Naturalesa dels dos aminoàcids involucrats
- ? Conformació dels segments que interaccionen de la cadena principal
→ valor dels angles planars dels carbonis alfa (segons la distribució bimodal)
- ? Orientació mútua dels dos grups laterals en contacte:
 - Paral.lela
 - Antiparal.lela
 - intermedia

Table 2. Contact potential for Lys-Glu interactions

	P	M	A
CC	-0.9	-0.4	0.9
EE	-1.1	-0.4	0.6
CE	-0.2	0.1	0.8
EC	-0.2	0.0	0.8

POTENCIALS ESTADÍSTICS

- ? Representen la tendència diferencial de residus a estar a prop o lluny en l'espai.
- ? S'utilitzen principalment per avaluar l'estabilitat de proteïnes ja plegades.

MÈTODES

? FOLDING THERMODYNAMICS OF PEPTIDES

- MOLECULAR DYNAMICS
- SOLVATION ENERGIES

? MONTE CARLO-METROPOLIS

? CONTACT MAPS

? MASSIVE PARALLEL COMPUTING

- REPLICA MD/MC COMPUTING

? ROSETTA

DINÀMICA MOLECULAR

? “Folding” → múltiples camins (no necessàriament el de menor energia)



Pas final involucra una ↓ nº intermediaris

? A partir de les poblacions relatives, càlcul de:

- l'energia lliure de “folding”
- diferència en l'energia lliure entre parells de conformacions

DINÀMICA MOLECULAR

? Stra. 3D d'un pèptid depèn de:

- Ambient
- Temperatura
- Pressió
- Solvent o molècules properes

**Distribució d'equilibris entre
diferents estructures**

**Si una conformació predomina sobre la resta en unes condicions
determinades, diem que el pèptid s'ha plegat**

DINÀMICA MOLECULAR

? Dades experimentals en els processos de determinació d'estructures, necessitat de:

- Ensamblatge
- Mitjana de temps entre diferents estructures

? Problema tècniques experimentals:

- NO permeten distingir entre $\neq$ stres. coexistents a la mostra durant la presa de dades.

DINÀMICA MOLECULAR

? MD

- Pèptid dins condicions determinades
- Simulació de l'evolució temporal del sistema



Distribució d'equilibris entre diferents conformacions



Anàlisi per mecànica estadística de la termodinàmica del sistema

DINÀMICA MOLECULAR

- ? Mecànica estadística de la termodinàmica del sistema

Probabilitat de trobar el pèptid amb una determinada conformació

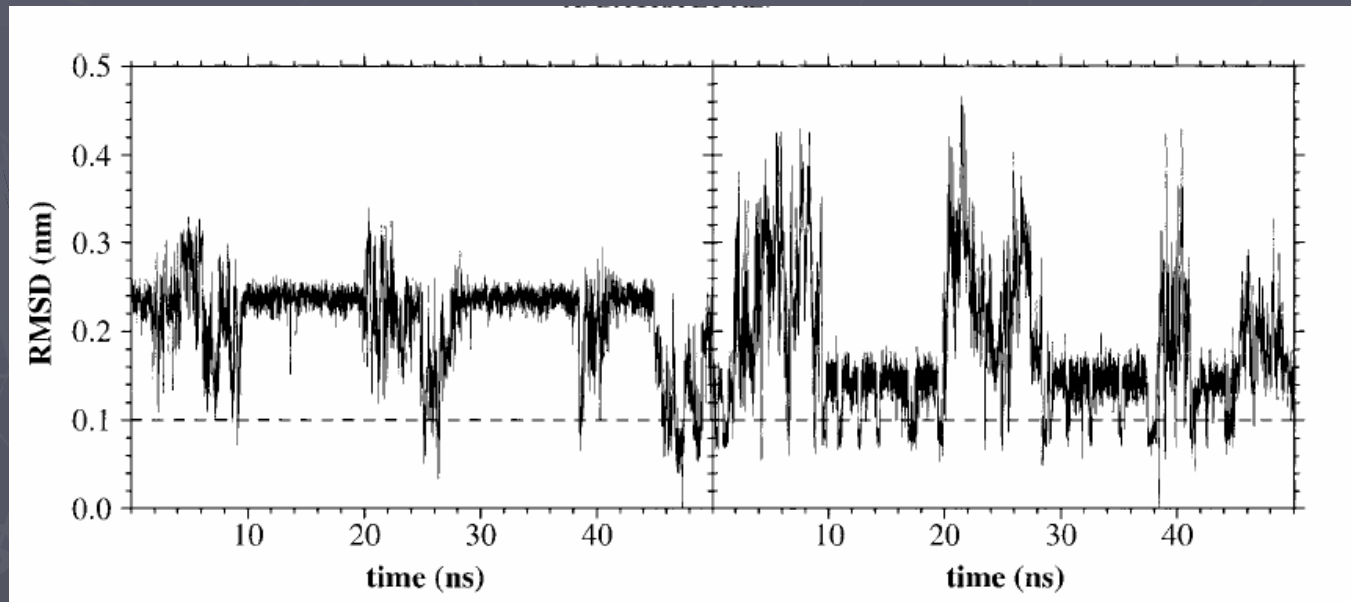
Estimació de l'energia lliure del plegament
Diferència en l'energia lliure de dos conformacions

Examinar camins “preferits” entre dos conformacions
Detectar intermediaris comuns (folding-unfolding)

DINÀMICA MOLECULAR

Conjunt de conformacions “unfolded” → petita proporció de l’espai conformacional accessible

El pèptid “**visita**” **repetidament** un petit nombre de conformacions durant una **escala temporal < temps de simulació**



Obtenció dels estadístics necessaris per valorar processos interconversió entre diferents conformacions

DINÀMICA MOLECULAR

? ENERGIA LLIURE

$$\Delta G_{A \rightarrow B} = - K_B T \ln \frac{P_B}{P_A}$$

? RATIOS DE TRANSICIÓ

$TR_{A \rightarrow B}$  Mitjana $\frac{\text{Vegades } A \rightarrow B}{\text{ns}}$ (interval de temps de 0,01 ns)

$TR_{A \leftrightarrow B}$  Mitjana $TR_{A \rightarrow B}$ i $TR_{B \rightarrow A}$

ENERGIES DE SOLVATACIÓ

? Energia lliure de solvatació

$$\Delta G_{\text{sol}} = \Delta G_{\text{ele}} + \Delta G_{\text{ster}}$$

$$\Delta G_{\text{ster}} = \Delta G_{\text{cav}} + \Delta G_{\text{disp-rep}}$$

$$\Delta G_{\text{ster}} = \sum_i \gamma_i S A S_t$$

ENERGIES DE SOLVATACIÓ

? Energia lliure (electrostàtica)

$$DG_{\text{ele}} = \sum_0^1 l_s V = \frac{1}{2} [l^2]_0^1 s V = \frac{1}{2} s V$$

$$\Delta G_{\text{ele}} = \frac{1}{2} \left\langle \sum_{i,j} \frac{Q_i Q_j}{r_{ij}} \right\rangle$$

? Partició de l'energia lliure de solvatació en la contribució de grups

$$\Delta G_{\text{ele}} = \sum_{i=1}^N \Delta G_{\text{ele},i} = \sum_{i=1}^N \frac{1}{2} \left\langle \sum_{i=1}^N \sum_{\substack{S=1 \\ S \in i}} \sum_{s=1}^{ns} \frac{Q_i Q_s}{|r_s - r_i|} \right\rangle$$

MONTE CARLO - METROPOLIS

? Tractar el procés com en un Model de Markov



Pas d'un estat al següent → matriu de probabilitats de transició

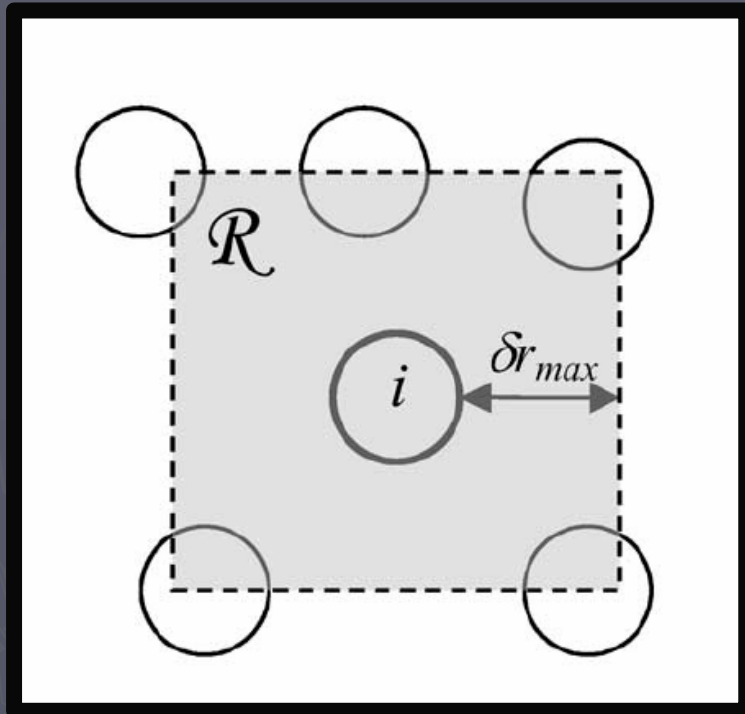
? Simulacions MC utilitzant l'energia del sistema



Criteri per avaluar l'acceptació/rebuig d'un pas de MC



MONTE CARLO - METROPOLIS



$$r_i^n = r_i^m + (2a_0 - 1)\delta r_{max}$$

$r_i^n \rightarrow$ localització final de la partícula i

$r_i^m \rightarrow$ localització inicial de la partícula i

$a_0 \rightarrow$ n° a l'atzar entre 0 i 1

$\delta r_{max} \rightarrow$ desplaçament màxim permès

L'estat n s'obté movent l'àtom i de l'estat m a qualsevol punt de la regió R amb una probabilitat uniforme.

MONTE CARLO - METROPOLIS

? **Avaluació d'acceptació/rebuig del desplaçament proposat**



Criteri d'acceptació → canvi en l'energia potencial del sistema

$$\delta V_{n,m} = \left(\sum_{j=1}^N v(r_{i,j}^n) - \sum_{j=1}^N v(r_{i,j}^m) \right)$$

Energia potencial

Es suma l'energia potencial de totes les partícules a una distància límit r_c en relació a la partícula i

MONTE CARLO - METROPOLIS

$dV_{n,m} < 0$ ® moviment $\rightarrow$, **acceptat**.

$dV_{n,m} > 0$ ® moviment - 

$$\frac{\rho_n}{\rho_m} = \frac{\exp(-\beta V_n) \exp(-\beta \delta V_{n,m})}{\exp(-\beta V_n)} = \exp(-\beta \delta V_{n,m})$$

Factor de Boltzman (diferència energètica)

$$\beta = 1 / k_B T$$

En aquest procediment, es genera un nombre a l'atzar (z ® entre 0 i 1),
i es compara amb $\exp(-\beta dV_{n,m})$.

MONTE CARLO - METROPOLIS

$dV_{n,m} < 0$ ® moviment -, **acceptat**.

$dV_{n,m} > 0$ ® moviment -

$\left\{ \begin{array}{l} \zeta < \exp(-bdV_{n,m}) \text{ ® } \mathbf{acceptació} \\ \zeta > \exp(-bdV_{n,m}) \text{ ® } \mathbf{rebuig del model proposat} \end{array} \right.$

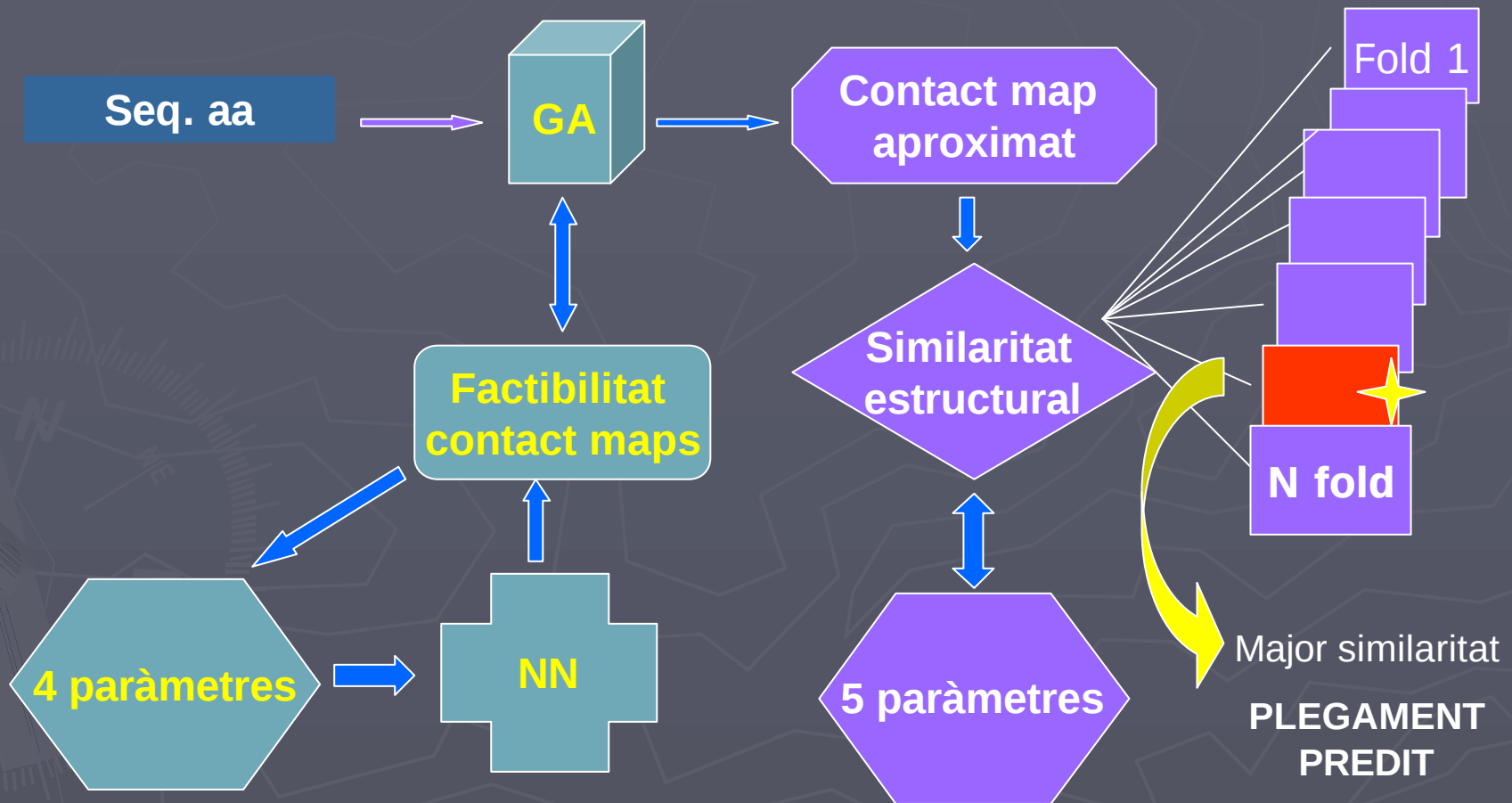
La partícula es manté en la seva posició original

El nou estat és el de **no-moviment**

Es continua el procés de modelatge seleccionant a l'atzar partícules successives pels desplaçaments proposats

CONTACT MAPS

- ? Contact Map → representació gràfica teòrica reduïda de prot. , que modela els contactes entre residus (a nivell local i global).

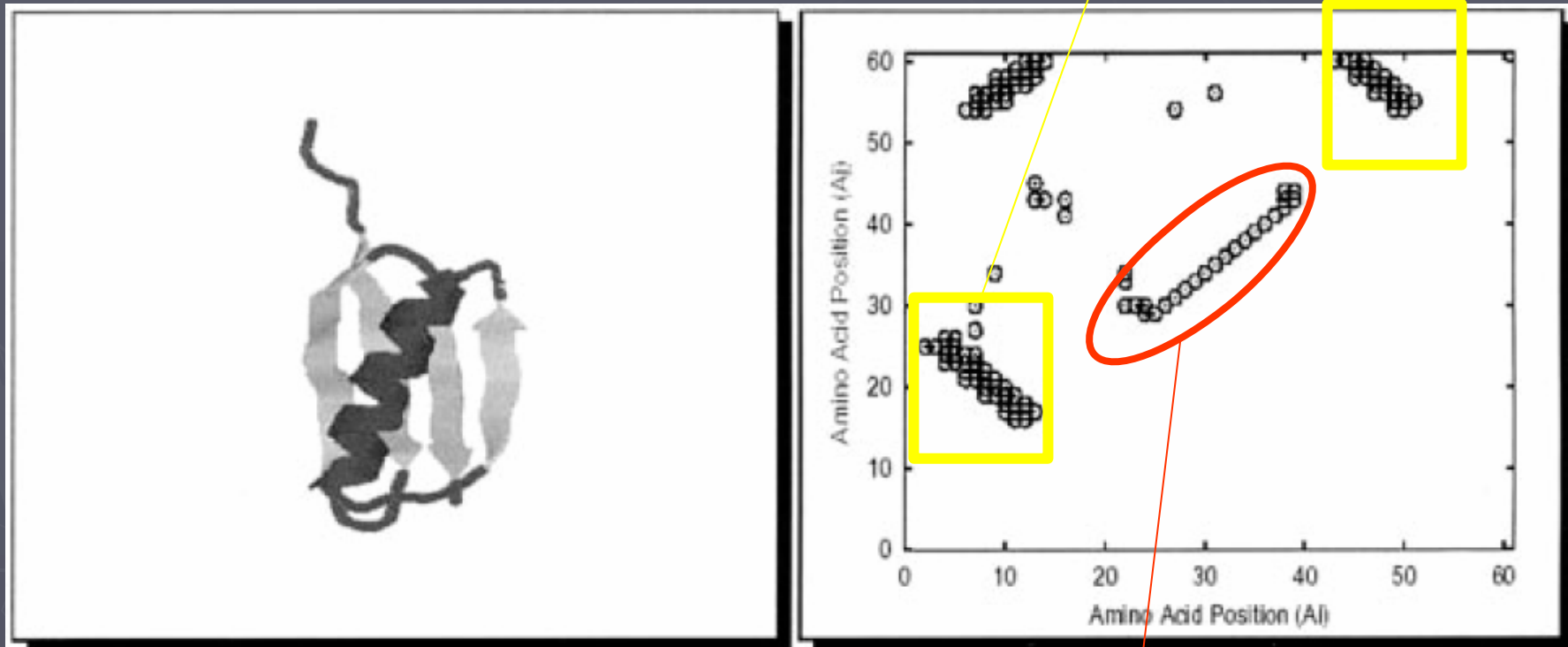


CONTACT MAPS

- ? Bons per predir stres. 2àries, però no tan bons per stres. 3àries.
- ? Identificació de “contact maps” factibles → 4 PARÀMETRES
 - ? Forces electrostàtiques
 - ? Distància a la seqüència
 - ? Hidrofobicitat
 - ? Graus de llibertat

CONTACT MAPS: exemple

? Proteïna 2igd



a-helix

MASSIVE PARALLEL COMPUTING

- ? **M simulacions independents** (des d'una condició inicial) analitzades en paral·lel
- ? 1a simulació que creua la **barrera d'energia lliure**
 - M vegades menys per creuar la barrera que el temps promig
- ? **Tornar** a començar M simulacions amb la nova localització després de la transició
- ? **Computacionalment econòmic** →
 - No gasta recursos de comunicació
 - Natural per a grups grans i heterogenis

REPLICA MD/ MC METHOD

MD: Molecular dynamics

MC: Monte Carlo-Metropolis

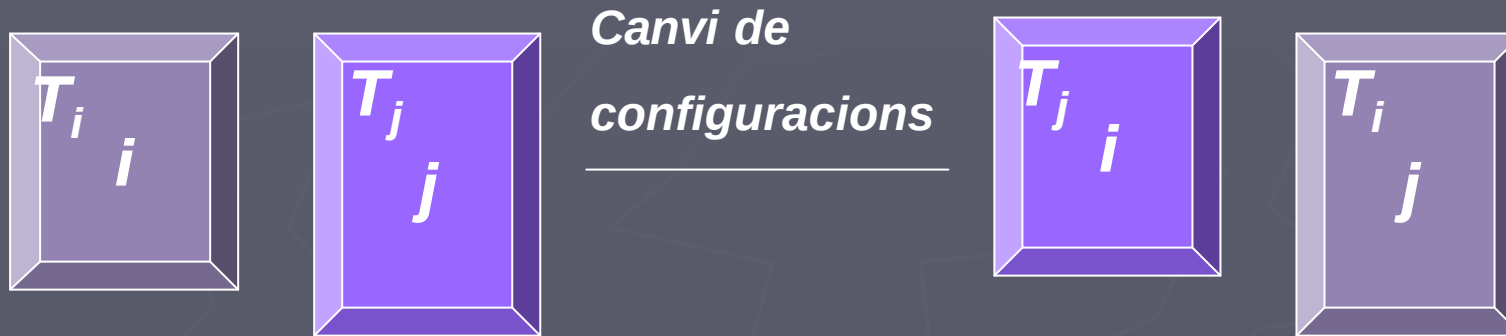
? Un nombre de **M rèpliques** distribuïdes sobre **M processadors**



? **Acoplament de rèpliques: canvi de T^a** (procediment MC)

REPLICA MD/ MC METHOD

A intervals fixes de temps:



Probabilitat del canvi de configuració:

$$W(X) \cdot w(X, X') = W(X') \cdot w(X', X)$$

$W(X)$ = factor de pes per l'estat X . Ve donat pel producte dels factors de Boltzman de cadascuna de les rèpliques.

$w(X, X')$ = probabilitat de transició per canviar del sistema X al sistema X'

REPLICA MD/ MC METHOD

Probabilitat del canvi de configuració:

$$W(X) \cdot w(X, X') = W(X') \cdot w(X', X)$$

$$W(X) = \exp [-\sum_i^M \beta_i E(\bar{x}_i, \bar{p}_i)]$$

$$E = E_{\text{kin}} + U$$

$$\frac{w(X, X')}{w(X', X)} = \exp(-\Delta)$$

$$\Delta = (\beta_i - \beta_j) \cdot (U_i - U_j)$$

$$(\beta_i - \beta_j) = 1 / k_B T_{\text{eff}}$$

$$T_{\text{eff}} = \frac{T_i T_j}{|T_i - T_j|}$$

$$T_{\text{eff}} \sim 15000 \text{ K} \quad \text{per a} \begin{cases} T_i = 300 \text{ K} \\ T_i - T_j \sim 6 \text{ K} \end{cases}$$

MC

REPLICA MD/ MC METHOD

- ? Probabilitats de transició: implementades utilitzant el criteri MC
- ? Només intercanvi de rèpliques amb T^{es} properes
- ? Direcció de l'intercanvi de rèpliques properes triat a l'atzar

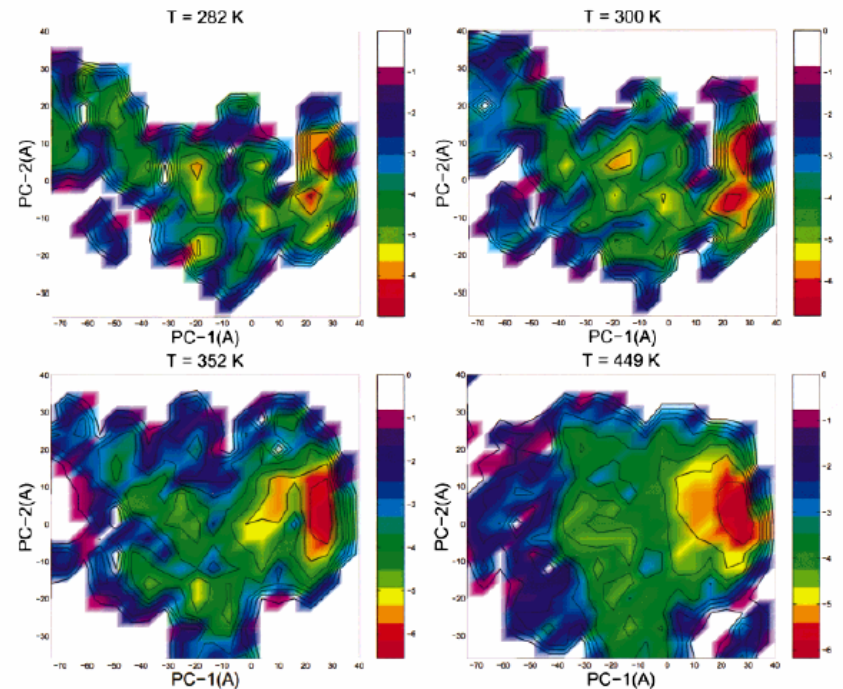
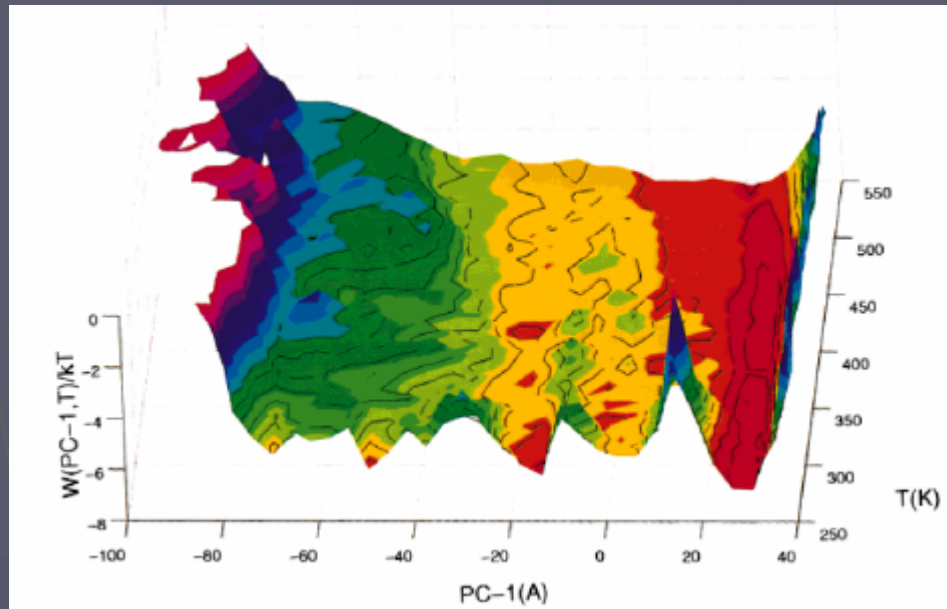
? **ANÀLISI DE COMPONENTS PRINCIPALS (PC):**

- Es calculen incloent juntes les fluctuacions de totes les rèpliques a totes les T^{es}
- Descriure les superfícies energètiques a cada T^{a}
- No s'inclouen tots els H

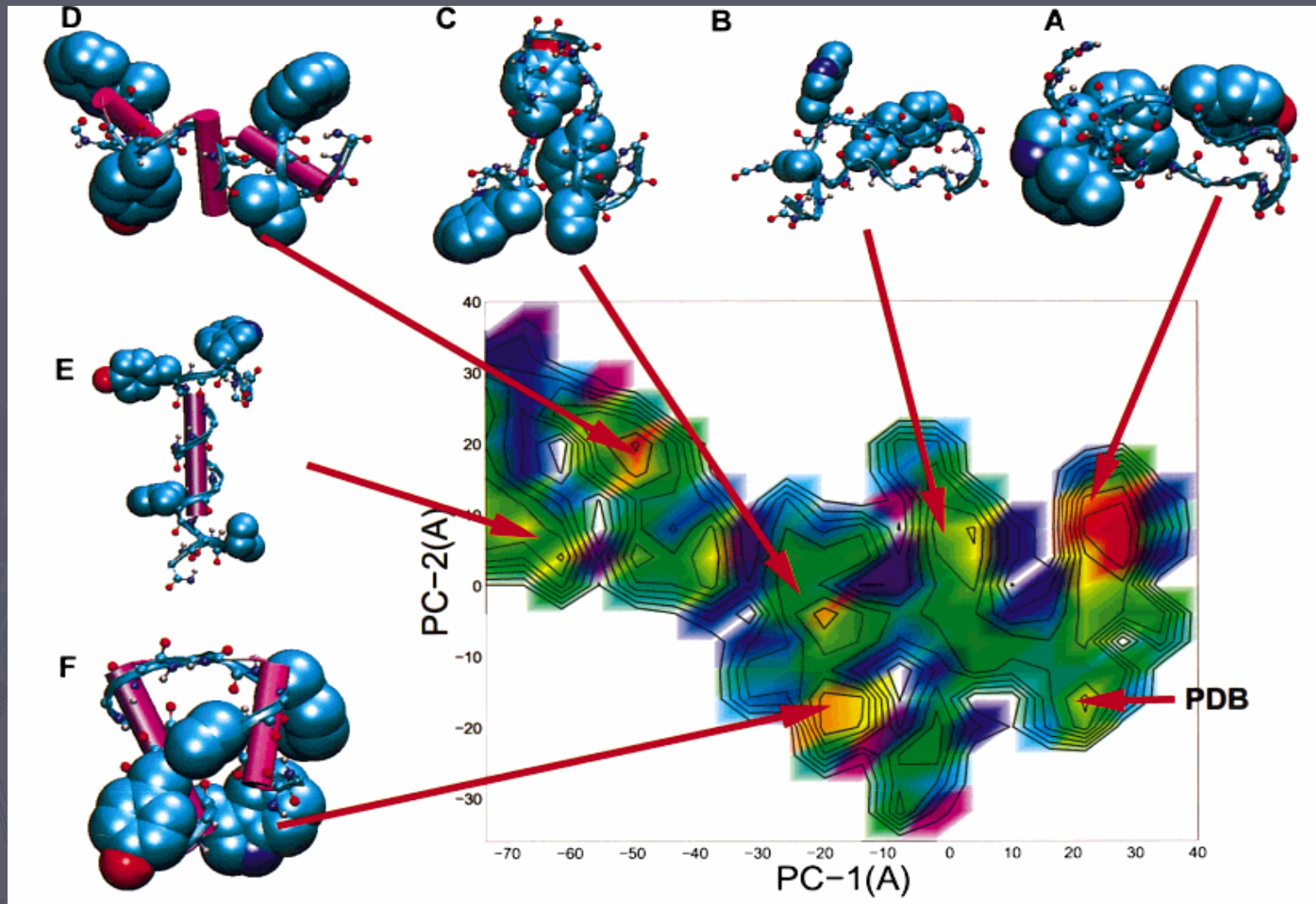
REPLICA MD/ MC METHOD: exemple

- ? Pèptid: **fragment C-terminal (41-56)** de la proteïna GB1
- ? **Solució acuosa**
- ? Forma estats “**glòbul desfet**” amb estructures similars als **β -hairpin**
a $T^{\text{es}} \sim 330 \text{ K}$
- ? **Múltiples mínims energètics** locals (altres estructures)
- ? $A \rightarrow T^{\text{es}} \rightarrow \uparrow n^{\circ} \text{ ponts d'H}$
- ? $T \sim 300^{\circ} \text{ K} \rightarrow$ plegament per formació de **regions hidrofòbiques**

REPLICA MD/ MC METHOD: exemple



REPLICA MD/ MC METHOD: exemple



MÈTODE ROSETTA

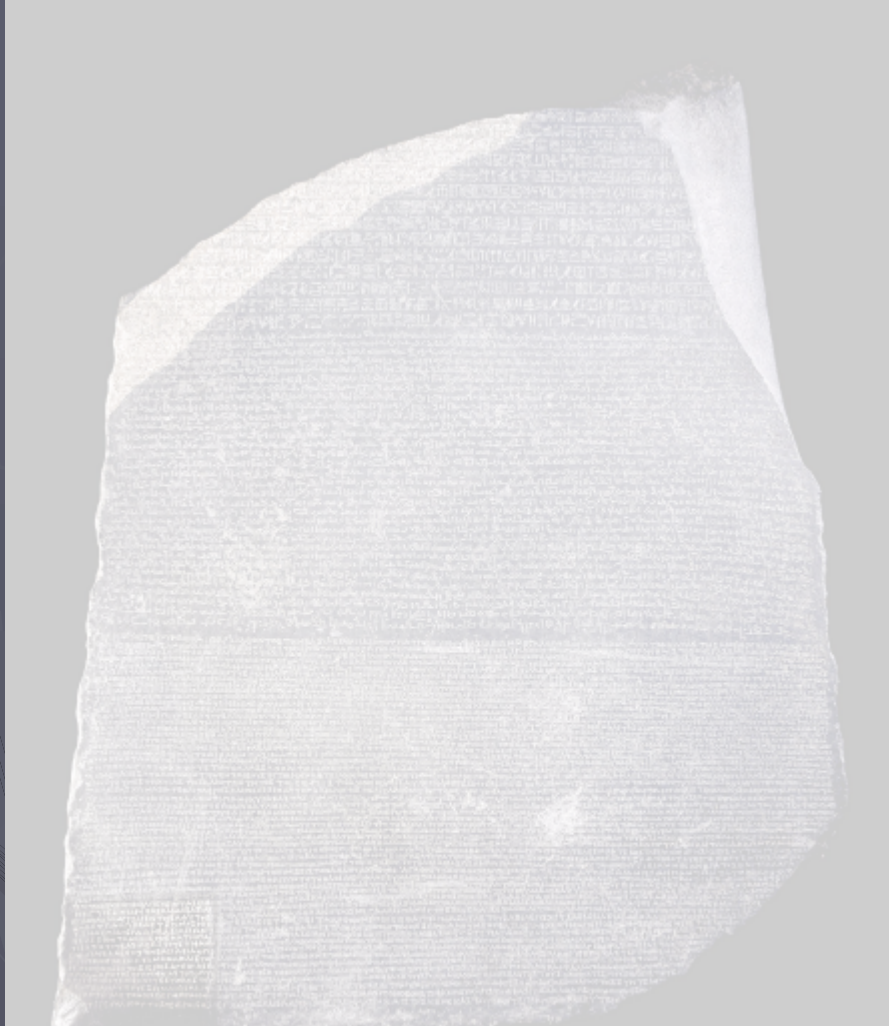
Manca de $F(x)$ de potencial prou acurades



Dificultat dels models *ab initio* per generar models raonables (coherents) !

Possible solució: limitar les possibles conformacions espacials a les compatibles per cada subsegment de la seqüència

MÈTODE ROSETTA



Kim Simons, David **Baker**,
Ingo Rudzinski and Charles
Kooperberg
(1997)

HMMSTR/Rosetta server:

<http://www.bioinfo.rpi.edu/~bystrc/hmmstr/server.php>

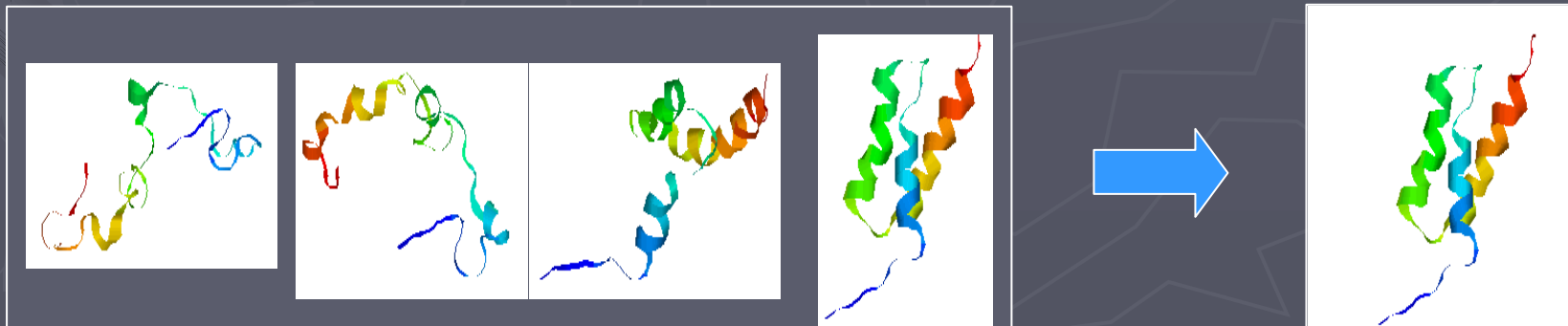
MÈTODE ROSETTA

➤ Primeres idees...

1. “En el plegament d’una proteïna, l’estructura local influeixia les conformacions possibles de segments curts de la seqüència”

2. “El plegament a l’estructura nativa es produeix quan aquests segments locals adopten simultàniament orientacions relatives que minimitzen l’energia lliure de l’estructura global”

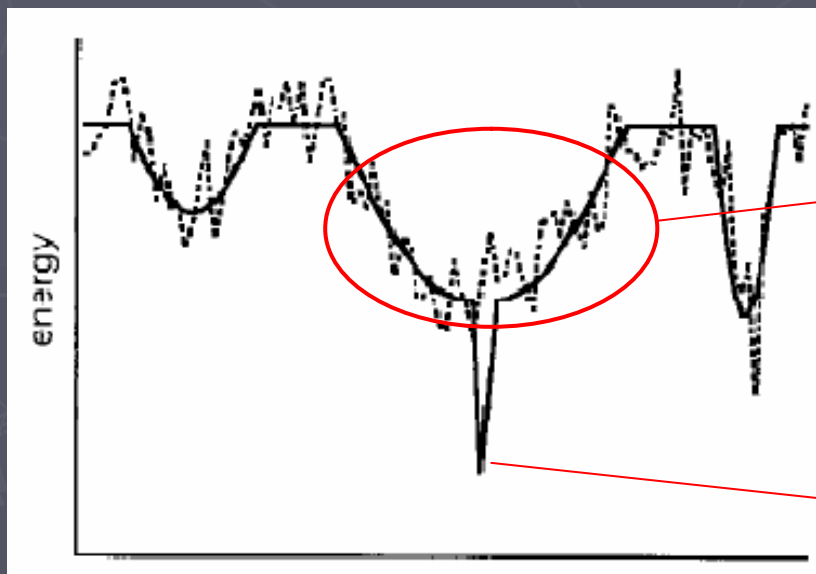
David Baker; JSB 134, 186-189 (2001)



MÈTODE ROSETTA

“...en el plegament, segments curts de la cadena oscilen entre diverses estructures locals, consistents amb la seva seqüència; el plegament a la estructura nativa es dona quan la orientació relativa dels diversos segments resulta en un mínim (o molt pròxim) global d'energia lliure.”

Simons et al., 1997



Near-native structures

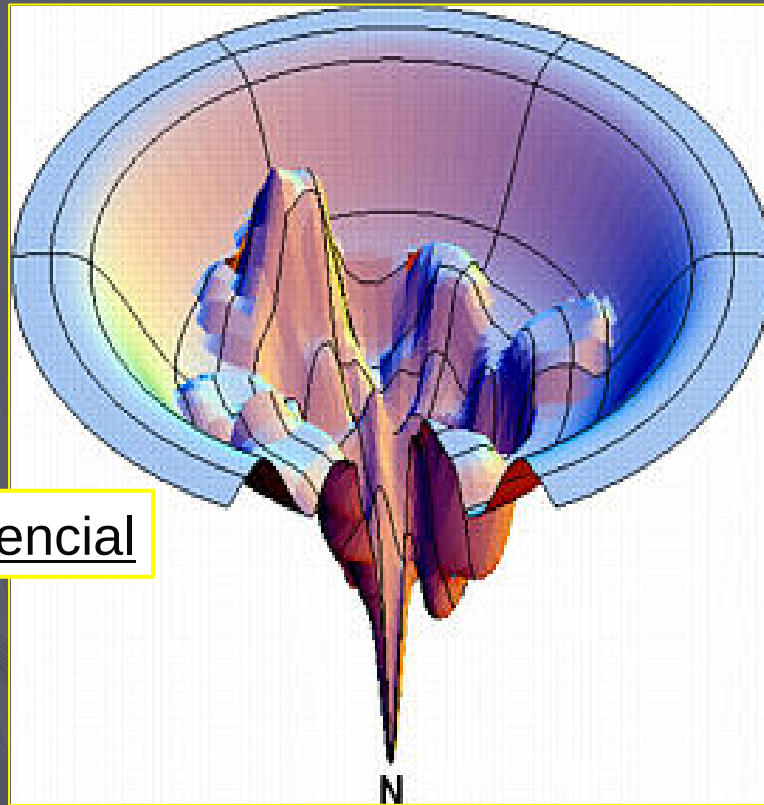
Estructura nativa

Plegament i Energia

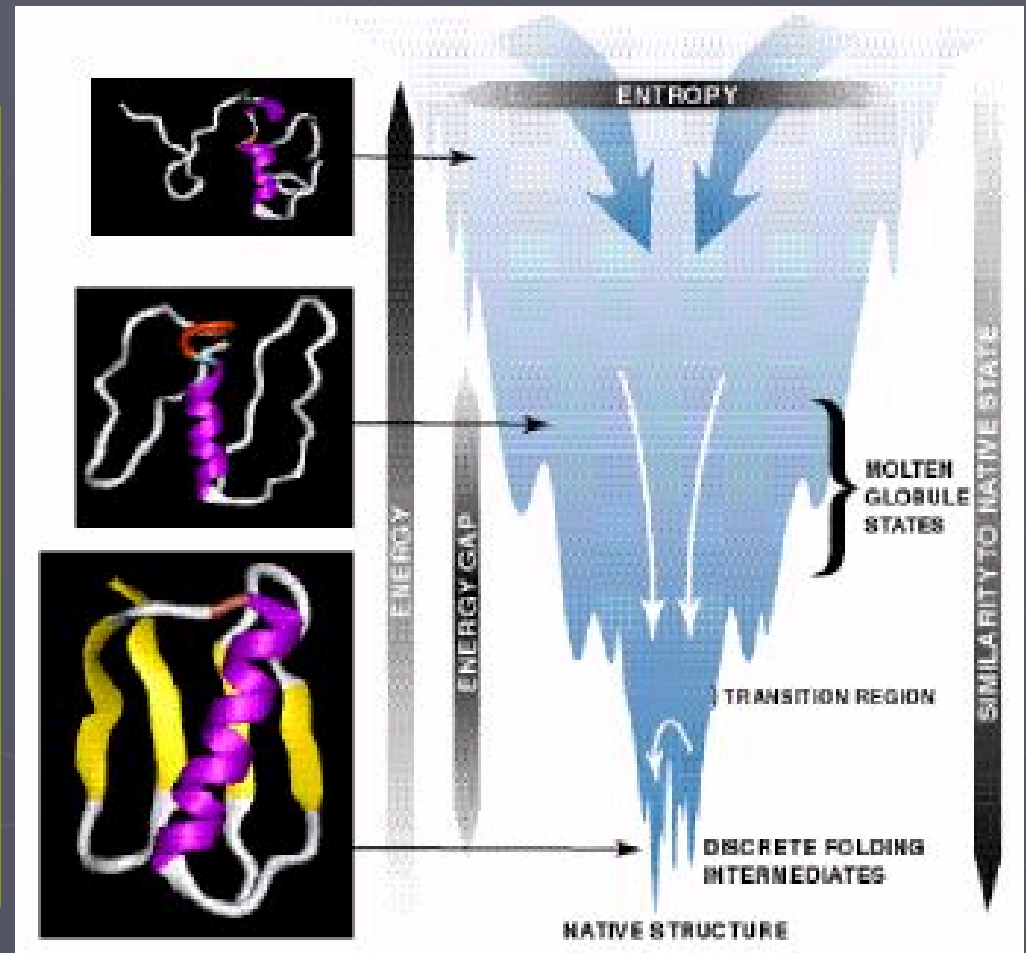
Espai configuracional

$\{\phi_1, \psi_1, \phi_2, \psi_2, \dots, \phi_n, \psi_n\}$

Potencial

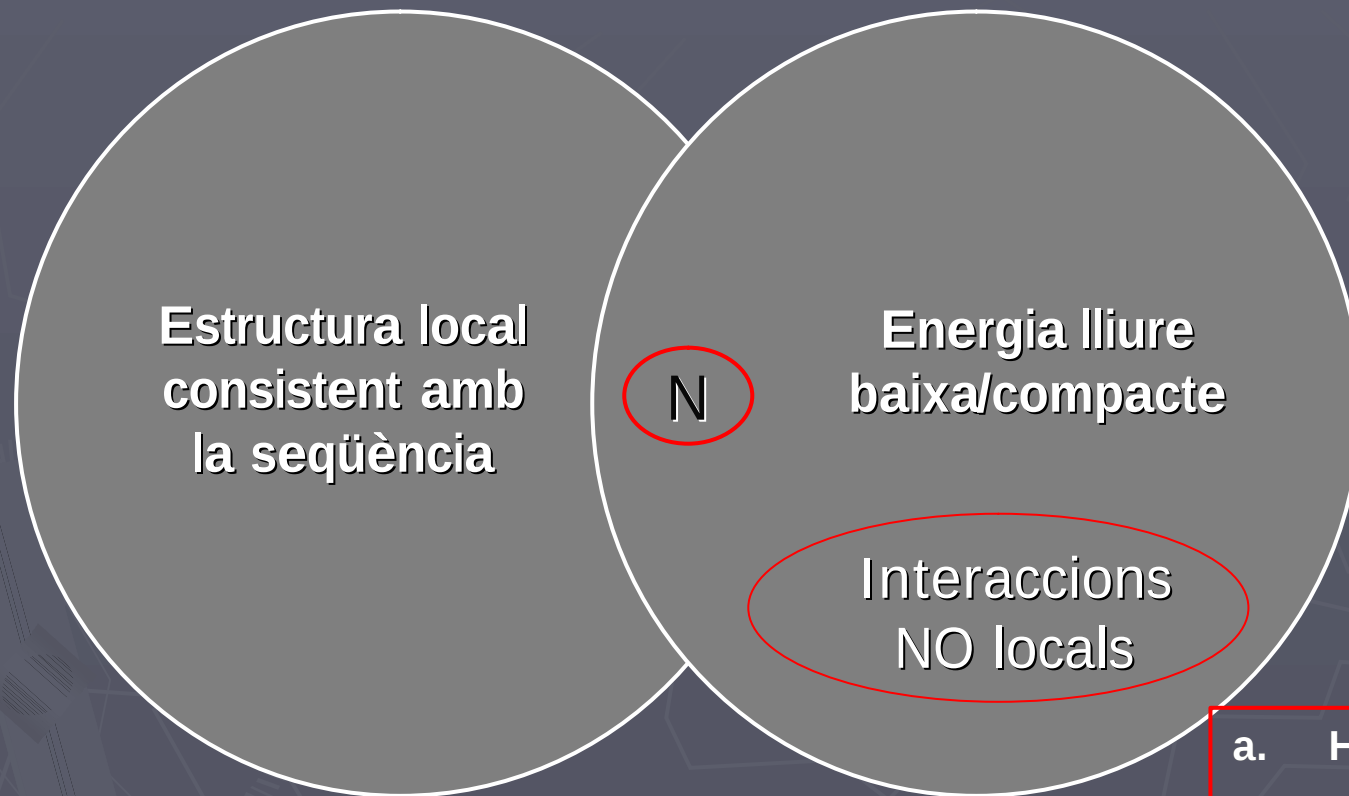


Nativa



Plegament i estructura

➤ Base conceptual del mètode Rosetta:



N = “Near-native structures”

- a. Hidrofobic burial
- b. Electrostatics
- c. Mainchain Hydrogen-bonding
- d. Excluded volume

MÈTODE ROSETTA

➤ Problema:

Models físics → no reproducció acurada d'interaccions subtils (predisposició local).

➤ Per tant...

“Assumim que la distribució de conformacions per cada segment curt de cadena polipeptídica és raonablement aproximada a la distribució de conformacions del mateix segment (i seqüències relacionades) en proteïnes d'estructura coneguda”.

MÈTODE ROSETTA

➤ PASOS:

- a) Crear una llibreria de “veïns estructurals”.
- b) Crear una colecció d'estructures conformacionals.
- c) **Filtrar**: descartem conformacions “non-native-like” (si pots...).
- d) **Clustering**; generem models amb les conformacions restants.

MÈTODE ROSETTA

a) LLIBRERIA VEÏNS ESTRUCTURALS

Segment/finestra de **9 residus [+ 3]** (millor correlació seqüència-estructura).

Generar freqüències de distribució pels 20 aminoàcids en segments de 9 residus i en seqüències de proteïnes d'estructura coneguda (**PDBselect_25**) **No homòlogues**.



“Distància” entre cada posició en el target i en les proteïnes d'estructura coneguda
(diferència mitjana de les freq. de distribució)

selecció dels **25-200** “veïns estructurals”

Versions actuals de “**Rosetta**” incorporen prediccions d'estructura secundària (**PSI-PRED,PHD,DSC**) per la selecció d'aquests veïns.

MÈTODE ROSETTA

b) COL·LECCIÓ D'ESTRUCTURES CONFORMACIONALS

➤ Inici amb la seqüència **primària** (cadena oberta).

➤ **Cicles d'estructuració:** n^o determinat (i.e. 1000) de canvis (i.e. angles de torsió).

- 1 canvi → canvia l'estructura en una posició random per l'estructura d'un dels 25 veïns estructurals.

MÈTODE ROSETTA

- Conformació avaluada (score). Pot acceptar-se o descartar-se.

SCORING FUNCTION!!



$$P(\text{structure}|\text{sequence}) = \frac{P(\text{sequence}|\text{structure}) * P(\text{structure})}{P(\text{sequence})}$$

$P(\text{structure}|\text{sequence}) \propto P(\text{sequence}|\text{structure}) * P(\text{structure})$;

- Terme per contribució **vdW** en l'score: evita la superposició d'àtoms.

MÈTODE ROSETTA

- Simulació **Metropoli Monte Carlo** (conformacions descartades no)

Com trobem un mínim d'energia lliure?

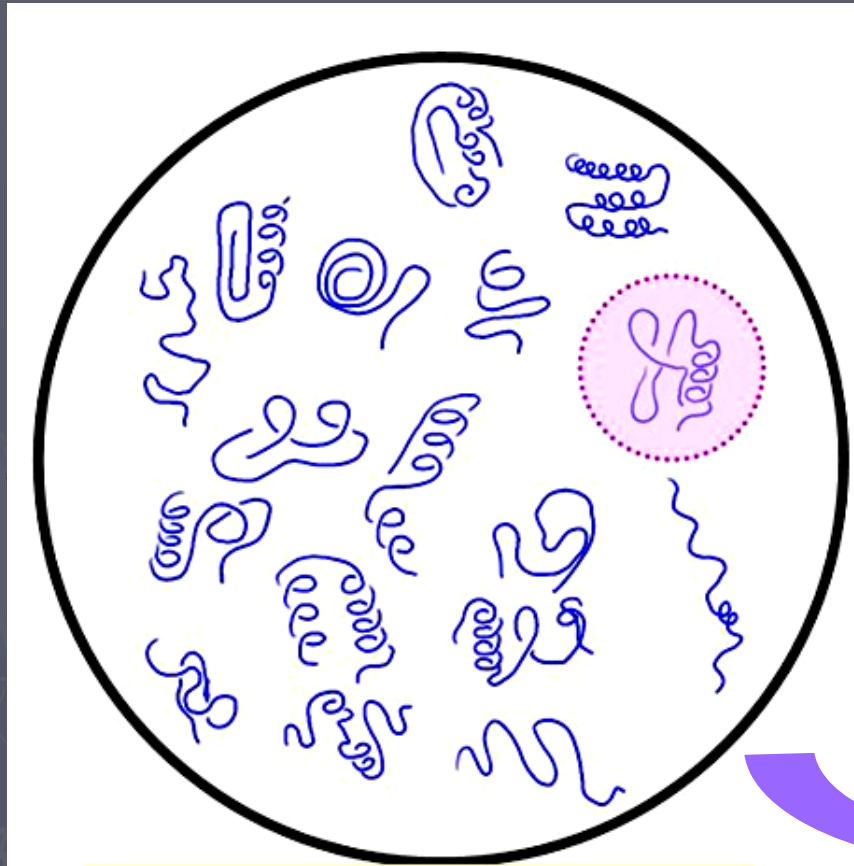
- Criteri per acceptar un canvi random=Energia lliure
- Si E augmenta: $P = (-\Delta E/T)$; T és l'energia artificial del procés

- Es genera un “decoy”.

- Repetim el procés per generar milers de “decoys”.

MÈTODE ROSETTA

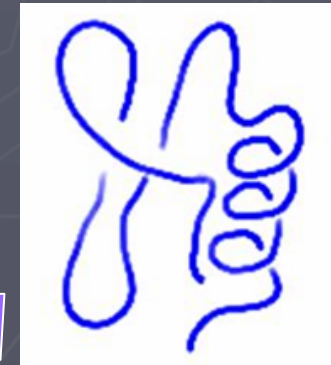
➤ Què és un “decoy”?



Decoys coherents

- Decoy: estructures generades, native i non-native-like.

- Discriminar



$F(x)$ d'energia acurades

MÈTODE ROSETTA

c) FILTRAR

- Propietats no consistents amb proteïnes conegudes.
- 2 filtres:
 - **CO mínim**
 - **cadena-β aparellades**
- **Contact Order:** Descarta formes amb **CO < 5è percentil** de les proteïnes d'estructura coneguda i mida similar.

MÈTODE ROSETTA

d) CLUSTERING

- Assumim: més conformacions **native-like** que **non-native like**.
- Distància entre decoys: **Ca RMSD**.
- Contem veïns per cada decoy dins un radi (**0,8-1nm**).
- Reduir radi seqüencialment, fins el "**decoy-model**".
- Mateix procés per cada cluster.
- Els 5 decoys amb millor score seran la base del modelatge final.

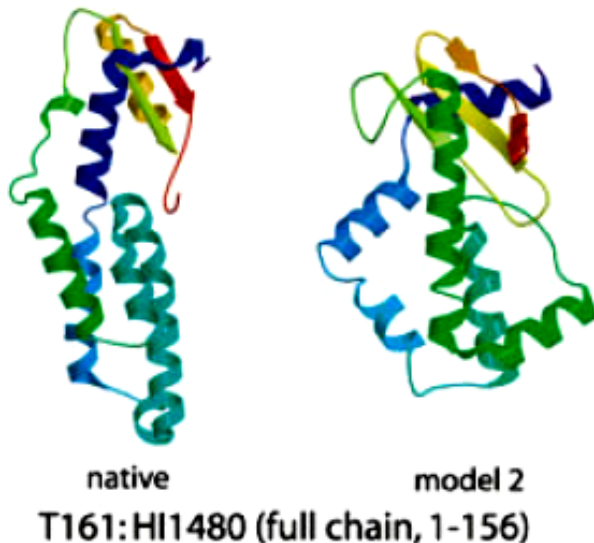
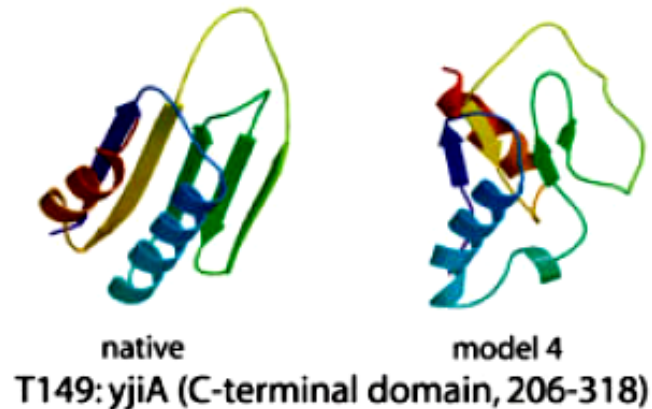
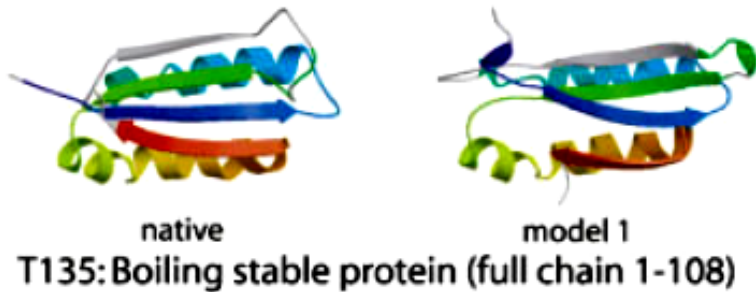
MÈTODE ROSETTA

➤ Com classifiquem els models?

- **Ca RMSD** : proteïnes petites.
- Classificació **SCOP** (superfamília): Mammoth Z-score.

Superfamília: inferències funcionals? Similaritat evolutiva?

Rosetta a CASP5



of residues with cRMS
below 4Å/6Å

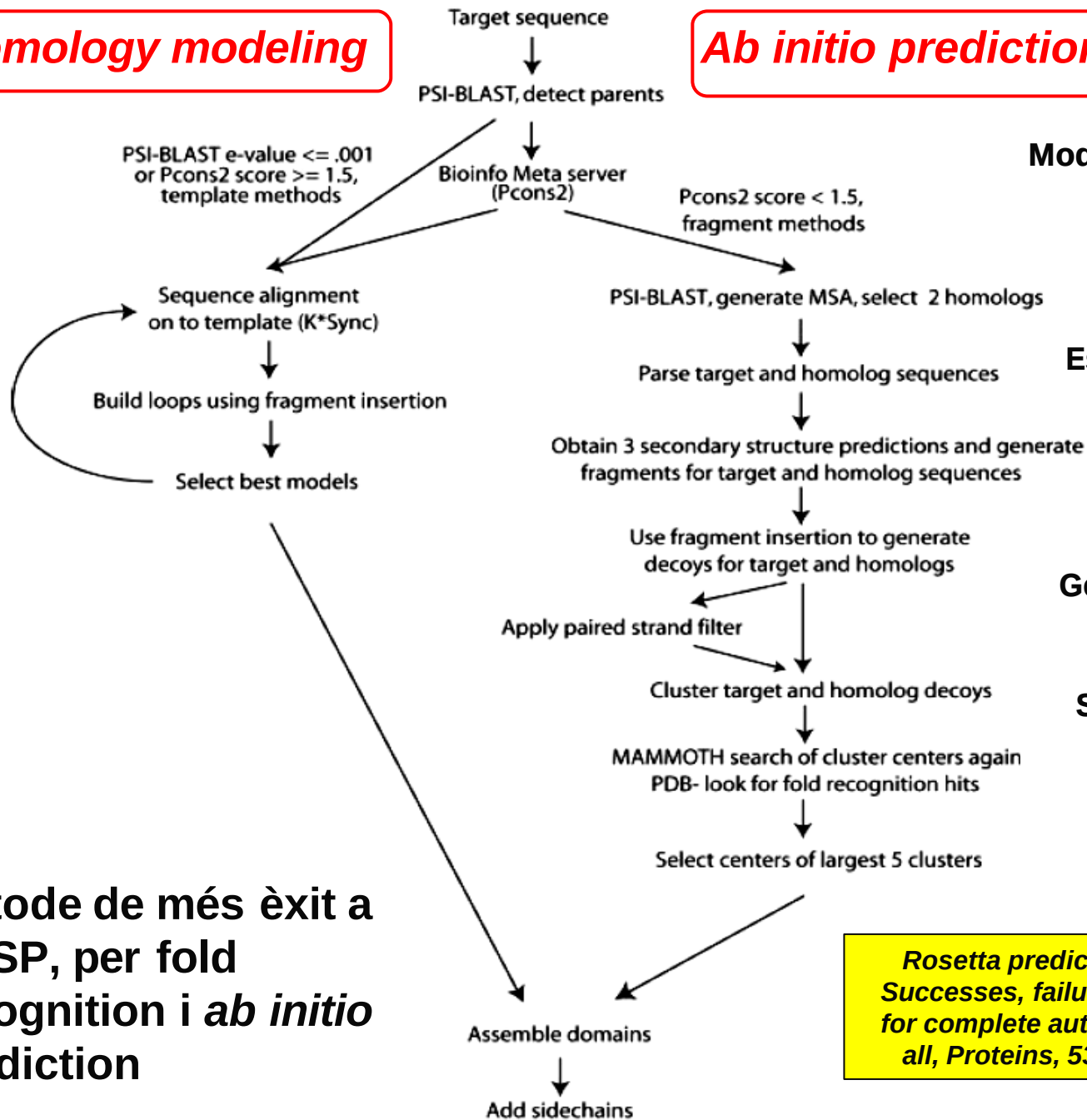
Name	Length	human	Automatic	Best decoy
T135	106	83/98	54/64	94/105
T149	116	52/71	44/62	76/92
T161	154	45/83	57/79	55/95

Rosetta predictions in CASP5:
Successes, failures, and prospect
for complete automation. Baker et
al, Proteins, 53:457-468 (2003)

ROSETTA PREDICTIONS IN CASP5

Homology modeling

Ab initio prediction



Modelar target i 2 homòlegs

Estructura 2aria

Generar decoys

Seleccionar 5 decoys per

Predicció

Mètode de més èxit a CASP, per fold recognition i *ab initio* prediction

Rosetta predictions in CASP5: Successes, failures, and prospect for complete automation. Baker et al, Proteins, 53:457-468 (2003)

MÈTODE ROSETTA

➤ Problemes:

- Excessiu nombre de conformacions (“decoys”).
- Baixa resolució dels models:
 - Millors models a CASP4: $\sim 6 - 10 \text{ \AA}$ RMSD $C\alpha$.
- Dificultat per distingir native de non-native. (score).

CASP

Critical Assessment of Techniques for Protein Structure Prediction

Protein Structure Prediction Center (<http://predictioncenter.llnl.gov/>)

➤ Estructures conegudes però no publicades. Disponibles al Maig; models pel Setembre; resultats al Desembre.

➤ Diverses categories:

- Homology-modelling, threading, *ab initio*,...

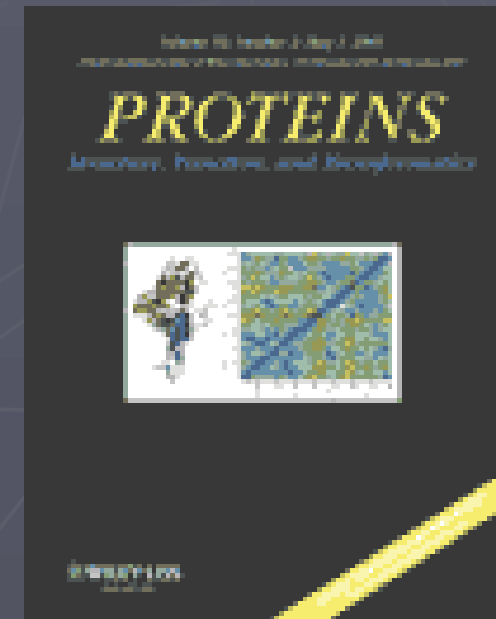
CASP

- Procés de “**blind prediction**”.
- Biannual.
- **Targets:** recull de dades experimentals (prèvia, o en procés de publicació). NMR, cristallografia,...
- Resultats es compilen i publiquen.
(<http://www3.interscience.wiley.com/cgi-bin/jissue/89510510>)
 - Altres categories
 - Participació oberta a tothom.

CASP

➤ FINALITAT:

- Test objectiu dels mètodes de predicció d'estructures.
- Referència de l'estat dels mètodes de predicció (progressos, errors i mancances).
- **Polèmica:** *temps i esforç perdut o assaig objectiu positiu i necessari?*



APLICACIONES

- ? Per què ens interessa ser capaços de predir l'estructura d'una proteïna *ab initio*.?
 - Estructura → Funció
 - Disseny racional de fàrmacs
 - Problemes per determinar experimentalment l'estructura de les proteïnes
 - Procés completament automatitzat
 - **Anotació del genoma: Identificació d'homòlegs llunyans**
 - **Modelatge per homologia**

APLICACIONES

Anotació del genoma

Tradicionalment basat en la similitud de seqüència

Inconvenients: {
inserció i /o extensió de dominis
permutacions circulars
intercanvi d'elements d'estructura secundària

El **mètode Rosetta** ha estat molt útil en aquest camp, en especial en la predicció estructural de **ORFs**.

APLICACIONES

- Modelatge per homologia

- ? Aquesta és principal aplicació dels mètodes *ab initio*.

- ? Plegament de zones de la proteïna problema (generalment els residus C i N-terminals) per les quals no hem trobat cap patró que ens doni informació.

BIBLIOGRAFIA

Bonneau R, Baker D. **AB INITIO PROTEIN STRUCTURE PREDICTION:** Progress and Prospects. *Annu. Rev. Biophys. Biomol. Struct.* 2001. 30:173–89

Kolinski A. **Protein modeling and structure prediction with a reduced representation.** *Acta Biochim Pol.* 2004;51(2):349-71. Review

Fábio L. Custódio, Hélio J.C. Barbosa and Laurent E. Dardenne
Investigation of the three-dimensional lattice HP protein folding model using a genetic algorithm. *Genetics and Molecular Biology*, 27, 4, 611-615 (2004)

Daura X, van Gunsteren WF, Mark AE. **Folding-unfolding thermodynamics of a beta-heptapeptide from equilibrium simulations.** *Proteins.* 1999 Feb 15;34(3):269-80.

Cavitación (Ing. Ariel R. Marchegiani). *Pequeñas centrales hidráulicas.*
fain.uncoma.edu.ar/centraleshidraulicas/archivos/PCH-%20CAVITACION.PDF

BIBLIOGRAFIA

Morreale A, de la Cruz X, Meyer T, Gelpi JL, Luque FJ, Orozco M. **Partition of protein solvation into group contributions from molecular dynamics simulations.** *Proteins*. 2005 Jan 1;58(1):101-9.

Chen JC, Kim AS. **Brownian Dynamics, Molecular Dynamics, and Monte Carlo modeling of colloidal systems.** *Adv Colloid Interface Sci*. 2004 Dec 31;112(1-3):159-73.

Gupta N, Mangal N, Biswas S. **Evolution and similarity evaluation of protein structures in contact map space.** *Proteins*. 2005 May 1;59(2):196-204

Shirts MR, Pande VS. **Mathematical analysis of coupled parallel simulations.** *Phys Rev Lett*. 2001 May 28;86(22):4983-7.

Garcia AE, Sabonmatsu KY. **Exploring the energy landscape of a beta hairpin in explicit solvent.** *Proteins*. 2001 Feb 15;42(3):345-54

BIBLIOGRAFIA

Simons K. T., Strauss C, Baker D.

Prospects for ab initio protein structural genomics; J Mol Biol. 2001 Mar 9;306(5):1191-9.

Bonneau R, Ruczinski I, Tsai J, Baker D. **Contact order and ab initio protein structure prediction**; Protein Sci. 2002 Aug;11(8):1937-44.

Richard Bonneau and David Baker. **Ab initio protein structure prediction: Progress and prospects**; Annu. Rev. Biomol. Struct. 2001. 30:173-89.

David Shortle, Kim T. Simons and David Baker; **Clustering of low-energy conformations near the native structures of small proteins**; Proc. Natl. Acad. Sci. USA, Vol. 95, pp 11158-11162.

BIBLIOGRAFIA

Tsaipai Wang and Jianzhong Shen. **Protein tertiary structure prediction: Basic concepts and applications**; https://www.the-agera.org/toni/cs401/spring03/Notes/g4/g4_pt1.pdf

David Baker et al.; **Prediction and design of macromolecular structures and interactions.**;
<http://www.hhmi.org/research/investigators/bakerd.html>

Bonneau R, Tsai J, Ruczinski I, Baker D. **Functional inferences from blind ab initio protein structure predictions.**

J Struct Biol. 2001 May-Jun;134(2-3):186-90. Review

PREGUNTES PEM

1.- En quins casos s'utilitza la predicció *ab initio*?

- a) en absència de seqüències homòlogues
- b) Sempre
- c) Només quan tenim seqüències homòlogues
- d) Per trobar nous plegaments

e) La a) i la d).

2.- Consideracions importants en la predicció *ab initio*:

- a) Temperatura
- b) Pressió
- c) Ambient
- d) Molècules properes

e) Totes les anteriors

PREGUNTES PEM

3.- Respecte el mètode Monte Carlo-Metropolis, assenyala la resposta correcta:

a) És un mètode estocàstic.

b) No té en compte la localització inicial dels àtoms.

c) No contempla els estats d'elevat moviment.

d) No tracta els processos com a models de Markov

e) Les simulacions no es realitzen en base a l'energia del sistema.

4.-Respecte les energies de solvatació, assenyala la resposta incorrecta:

a) Tenen en compte les càrregues electrostàtiques tant del solut com del solvent.

b) L'energia lliure de solvatació és el resultat del sumatori de les energies lliures electrostàtica i estèrica.

c) L'energia lliure estèrica no té en compte l'energia lliure de cavitació.

d) L'energia lliure estèrica té en compte la superfície atòmica accessible al solvent.

e) L'energia lliure estèrica té en compte l'energia lliure de dispersió-repulsió.

PREGUNTES PEM

5.- Els paràmetres a tenir en compte per l'identificació de “contact maps” factibles són:

a) Forces electrostàtiques i distància a la seqüència.

b) Graus de llibertat i hidrofobicitat

c) La a) i la b)

d) Ambient i energia lliure de solvatació

e) La a) i la d)

6.- Respecte el mètode Rosetta, assenjala l'opció falsa:

a) Creació de llibreria de veïns estructurals.

b) Creació de colecció d'estructures conformacionals.

c) Filtrar (descartar conformacions “non-native-like”).

d) Clustering (generar models amb les conformacions restants)

e) El model definitiu s'obté a partir d'una única estructura.

PREGUNTES PEM

7.- En relació als mètodes basats en dinàmica molecular, assenyala la incorrecta:

- a) El pèptid es troba dins condicions determinades.
- b) Es simula l'evolució temporal del sistema.
- c) Es crea la distribució d'equilibris entre diferents conformacions.
- d) S'analitza per mecànica estadística la termodinàmica del sistema.
- e) No es detecten estructures intermediàries.**

8.- Assenyala la opció falsa en relació al Mètode Rosetta...

- a) Permet obtenir models estructurals de seqüències sense seqüències homòlegs.
- b) No pot generar estructures d'alta resolució.**
- c) Pot detectar nous plegaments.
- d) Utilitza subseqüències de 9 residus obtenir conformacions locals.
- e) Obté models estructurals per seqüències de < 100 residus.

PREGUNTES PEM

9.- Sobre la predicció ab initio de les proteïnes:

- a) Actualment els models off-tallice són els mètodes d'elecció a l'hora de predir ab initio l'estructura terciària de les proteïnes.
- b) La hipòtesi termodinàmica en que es basa ens diu que la conformació que adopta una proteïna en plegar-se és aquella amb l'energia lliure global mínima.**
- c) En general els mètodes ab initio són molt eficaços per predir l'estructura de les all-beta proteins
- d) És necessari el coneixement de l'estructura secundària per poder utilitzar qualsevol dels mètodes termodinàmics.
- e) Totes són falses

10.-Predicció ab initio:

- a) El mètode Rosetta és el millor mètode de predicció *ab initio* que hi ha a la actualitat, segons el CASP5
- b) El mètode Montecarlo – Metropolis es un mètode de simulació basat en l'energia potencial del sistema
- c) Les dues anteriors són certes**
- d) Els mètodes ab initio prediuen l'estructura terciària de seqüències de qualsevol longitud
- e) Totes les anteriors són certes

PREGUNTES ASSAIG

1.- Característiques generals de la predicció *ab initio* de plegaments. Objectiu, principi en el que es basa. Requisits. Problemes

- Objectiu: Donada la seqüència d'aminoàcids d'una proteïna, deduir l'estructura del seu estat nadiu.

Principi→ **HIPÒTESI TERMODINÀMICA**: "l'estructura amb que una proteïna es plega és aquella amb l'energia lliure global més baixa

- Requisits:

- ? Trobar la funció que permeti calcular l'energia lliure d'una estructura→ que tingui en compte totes les energies que participen en el seu plegament
- ? Desenvolupar un protocol eficient per trobar l'estructura amb la mínima energia lliure
- ? Bon mètode de cerca per explorar l'espai conformational

? És difícil trobar la funció "perfecte": la que ens permeti diferenciar estructures correctes d'incorrectes

? Es necessita una enorme quantitat de recursos computacionals per resoldre una funció d'aquest tipus

PREGUNTES ASSAIG

2.- Enumera les aplicacions dels mètodes de predicció de plegaments *ab initio*.

Estructura → Funció

Disseny racional de fàrmacs

Problemes per determinar experimentalment l'estructura de les proteïnes

Procés completament automatitzat

Anotació del genoma: Identificació d'homòlegs llunyans

Modelatge per homologia

PREGUNTES ASSAIG

3.- Enumera els passos del mètode Rosetta

- a) Crear una llibreria de “veïns estructurals”.
- b) Crear una col·lecció d'estructures conformacionals.
- c) **Filtrar**: descartem conformacions “non-native-like” (si pots...).
- d) **Clustering**; generem models amb les conformacions restants.

4.-Problemes del mètode Rosetta

-Excessiu nombre de conformacions (“decoys”).

-Baixa resolució dels models:

*Millors models a CASP4: $\sim 6 - 10 \text{ \AA rmsd Ca.}$

-Dificultat per distingir native de non-native. (score).