Modélisation moléculaire par homologie

Dirk Stratmann (dirk.stratmann@upmc.fr)
www.impmc.upmc.fr/~stratmann

IMPMC, Sorbonne Université
BFA, Université Paris Cité

septembre 2024

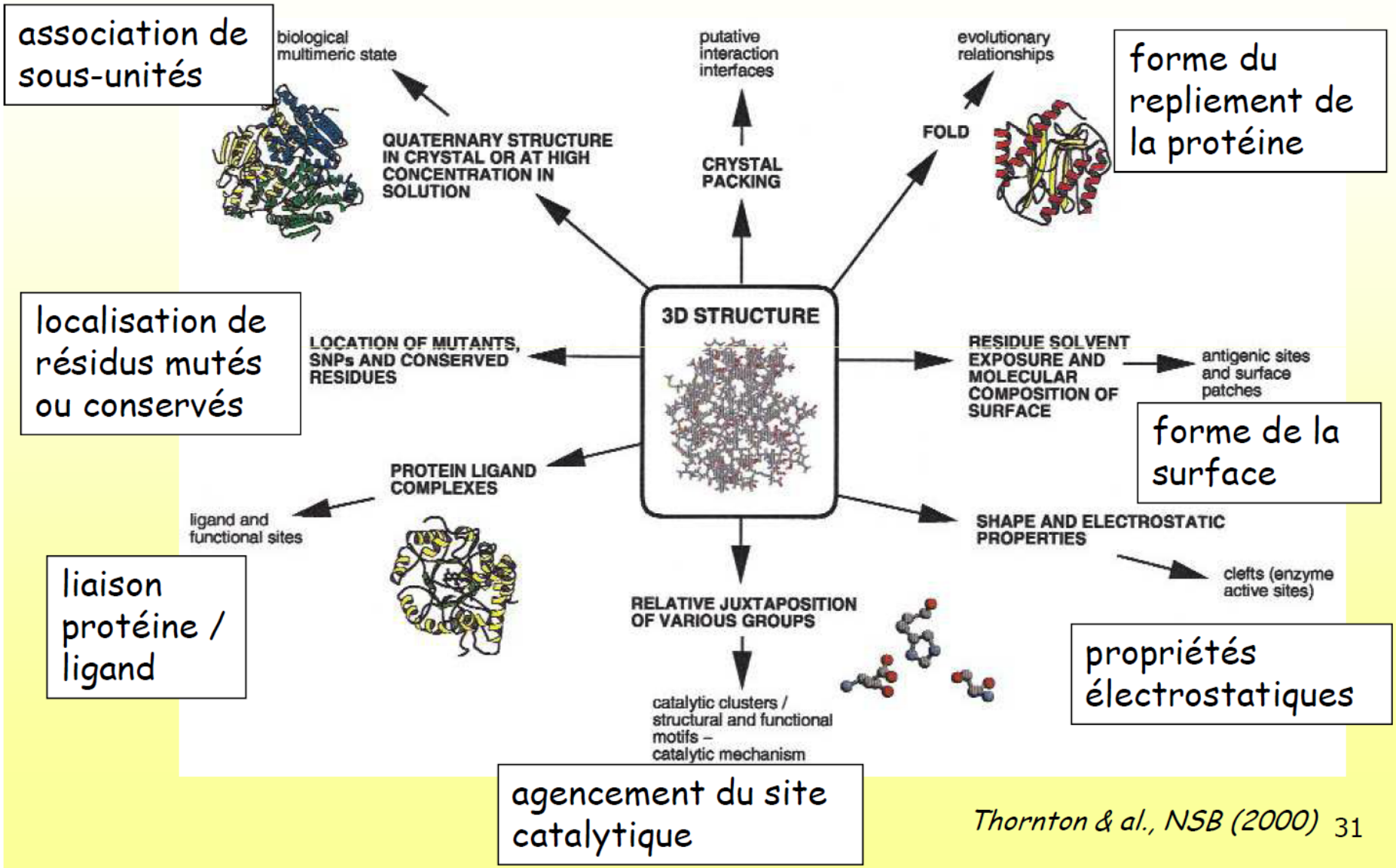
Plan

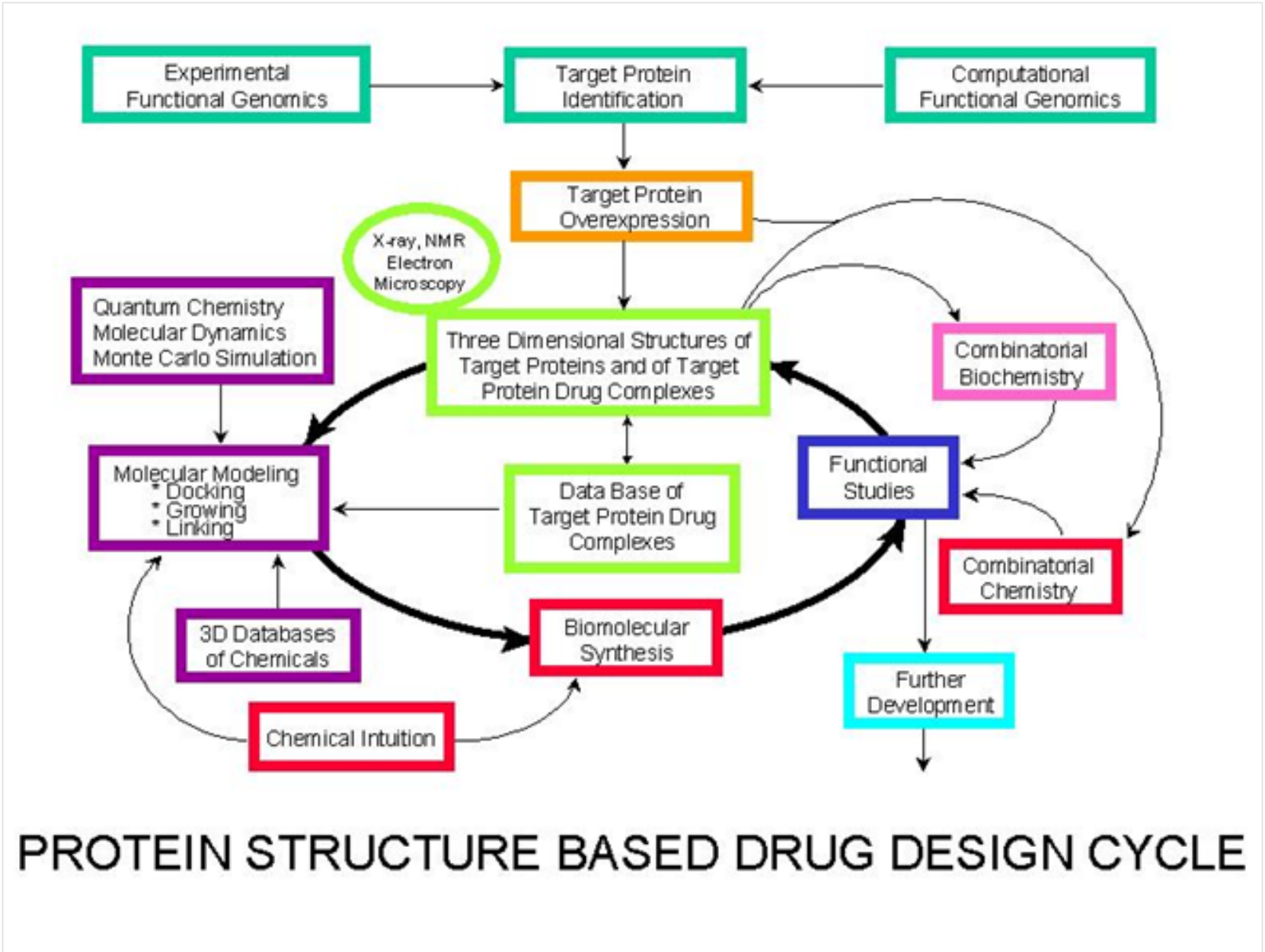
Introduction

Liens utiles

- Les PDF du cours seront déposés sur:
  `stratmann.fr/cours/homology_modeling`

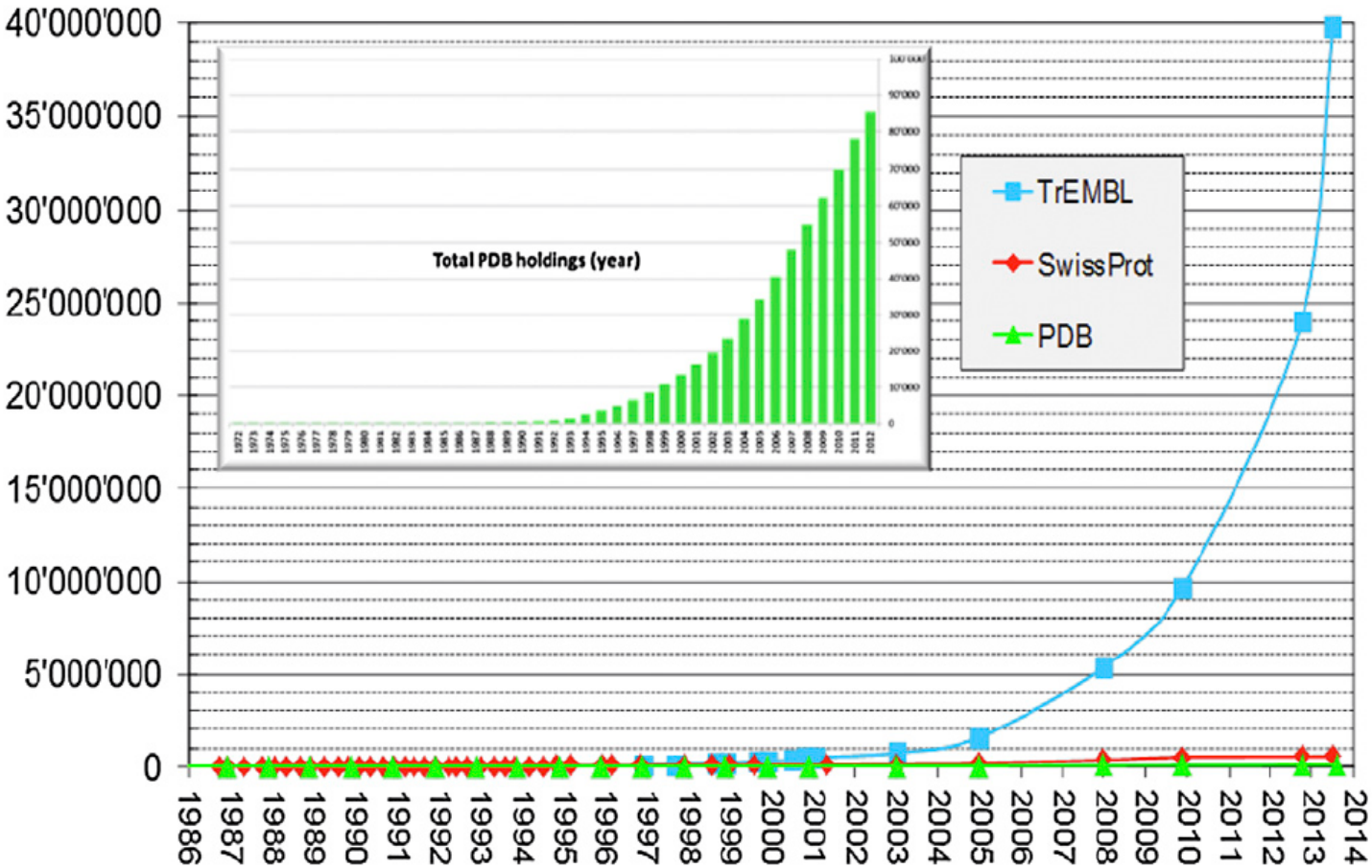
La structure 3D donne des informations sur la fonction biologique d'une protéine





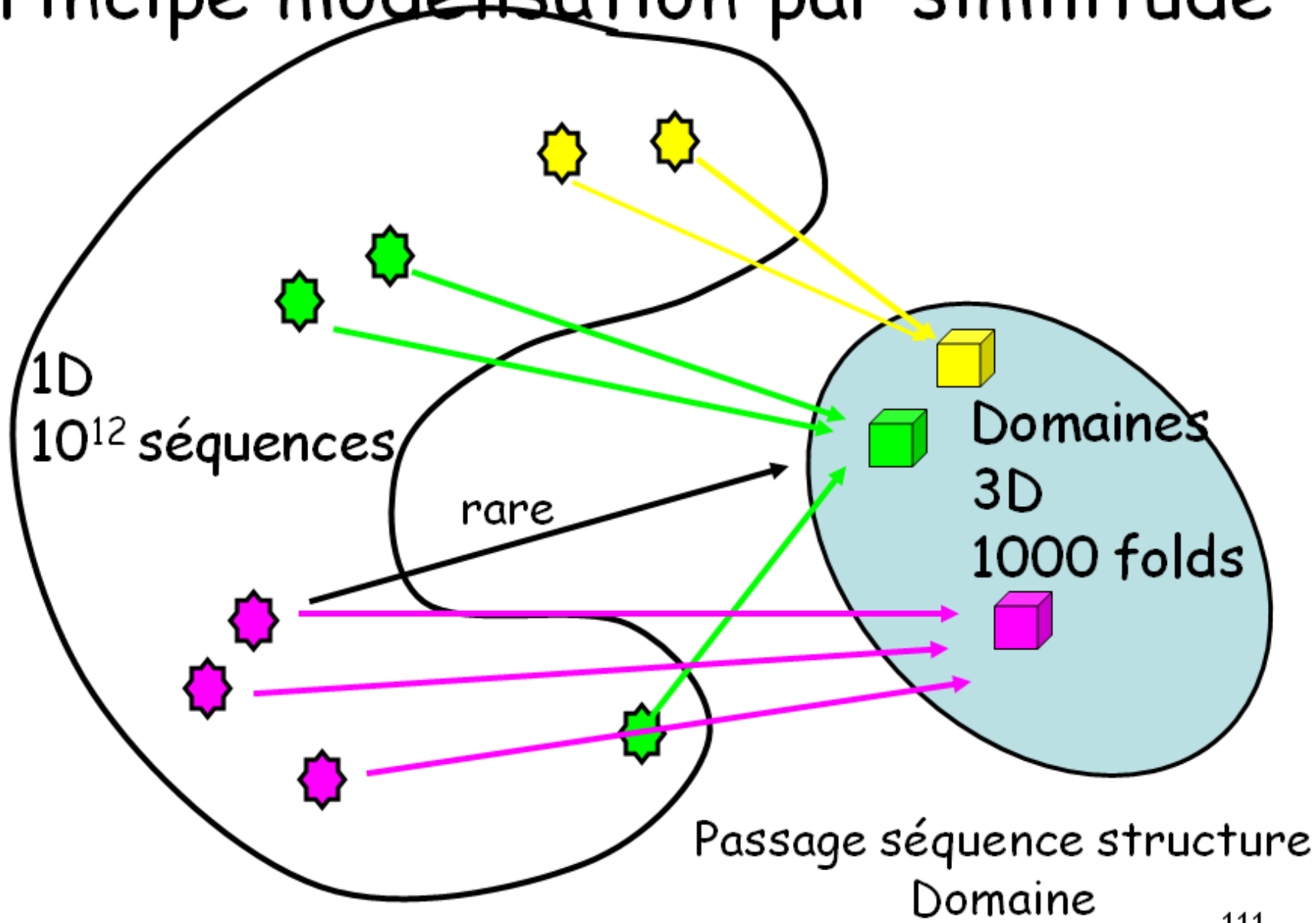
http://swift.cmbi.ru.nl/teach/B1SEM/Six_topics_FF_intro_seminar_B1SEM.ppt

Protein structure gap

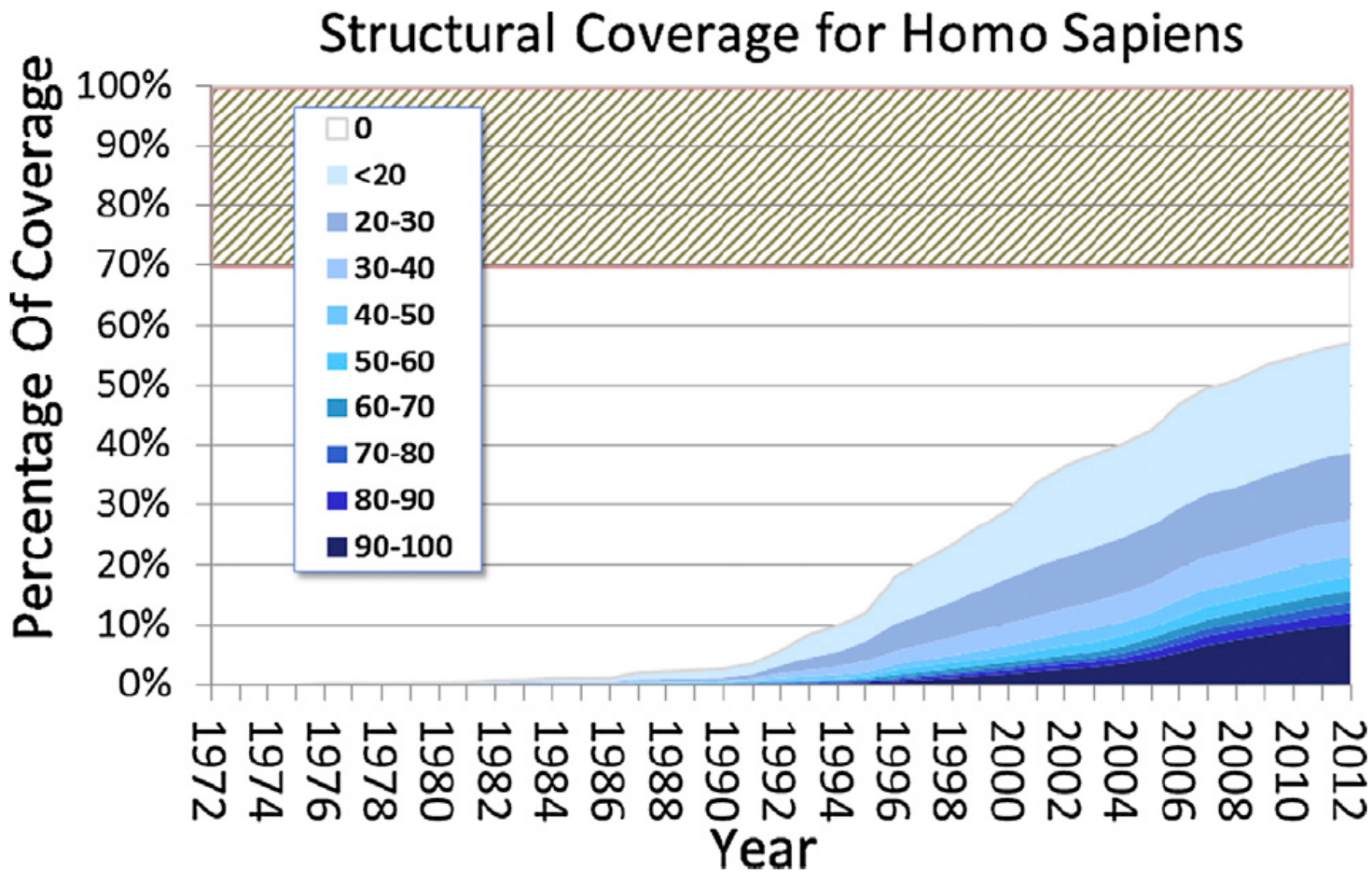


Torsten Schwede (Sept. 2013). In: *Structure* 21.9

Principe modélisation par similitude



Structural coverage



Torsten Schwede (Sept. 2013). In: *Structure* 21.9

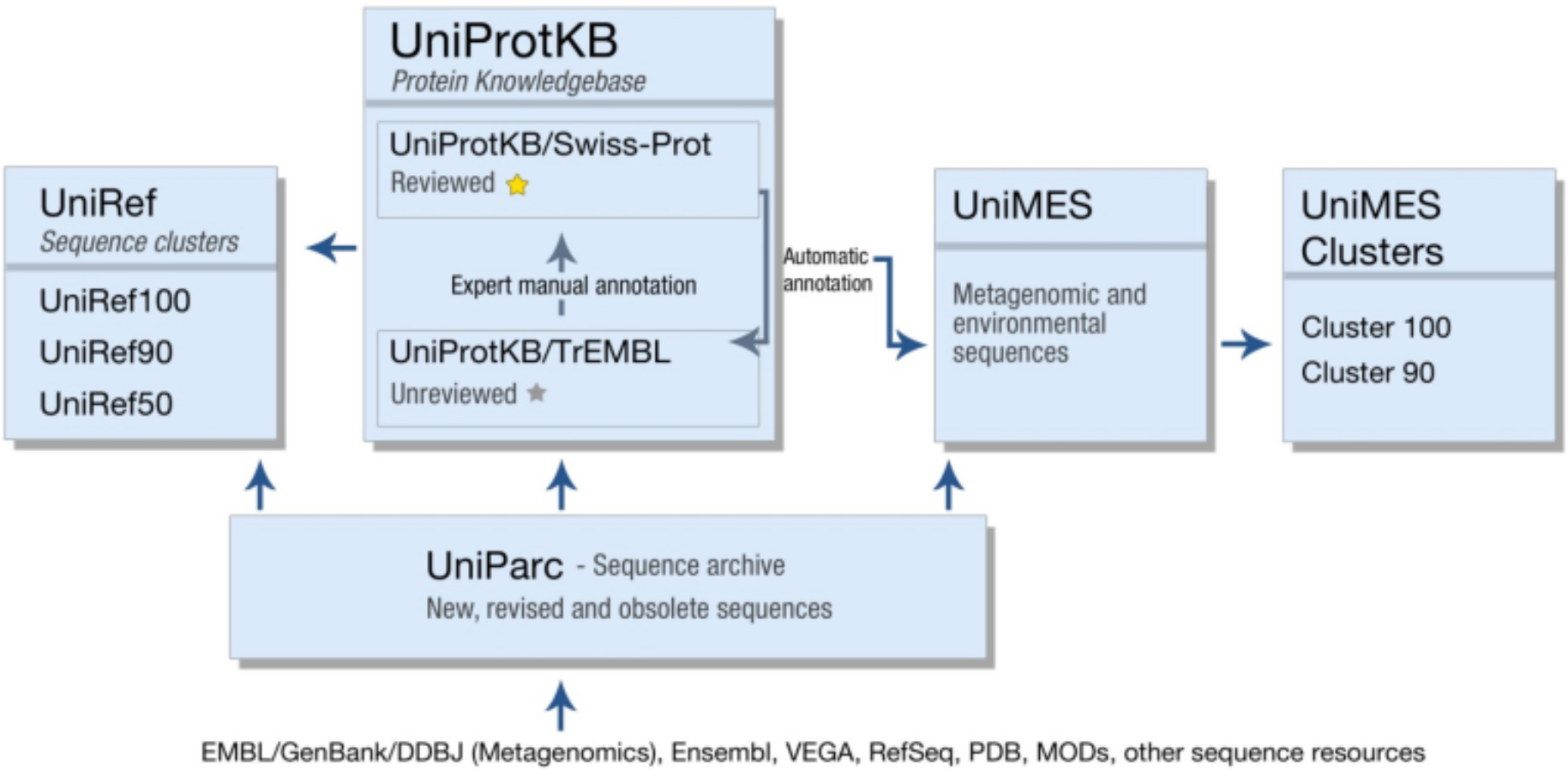
2 Alignement de séquences

- Alignement de séquences - Introduction
- Alignement en pratique
- Alignement multiple

Alignement de séquences - Introduction

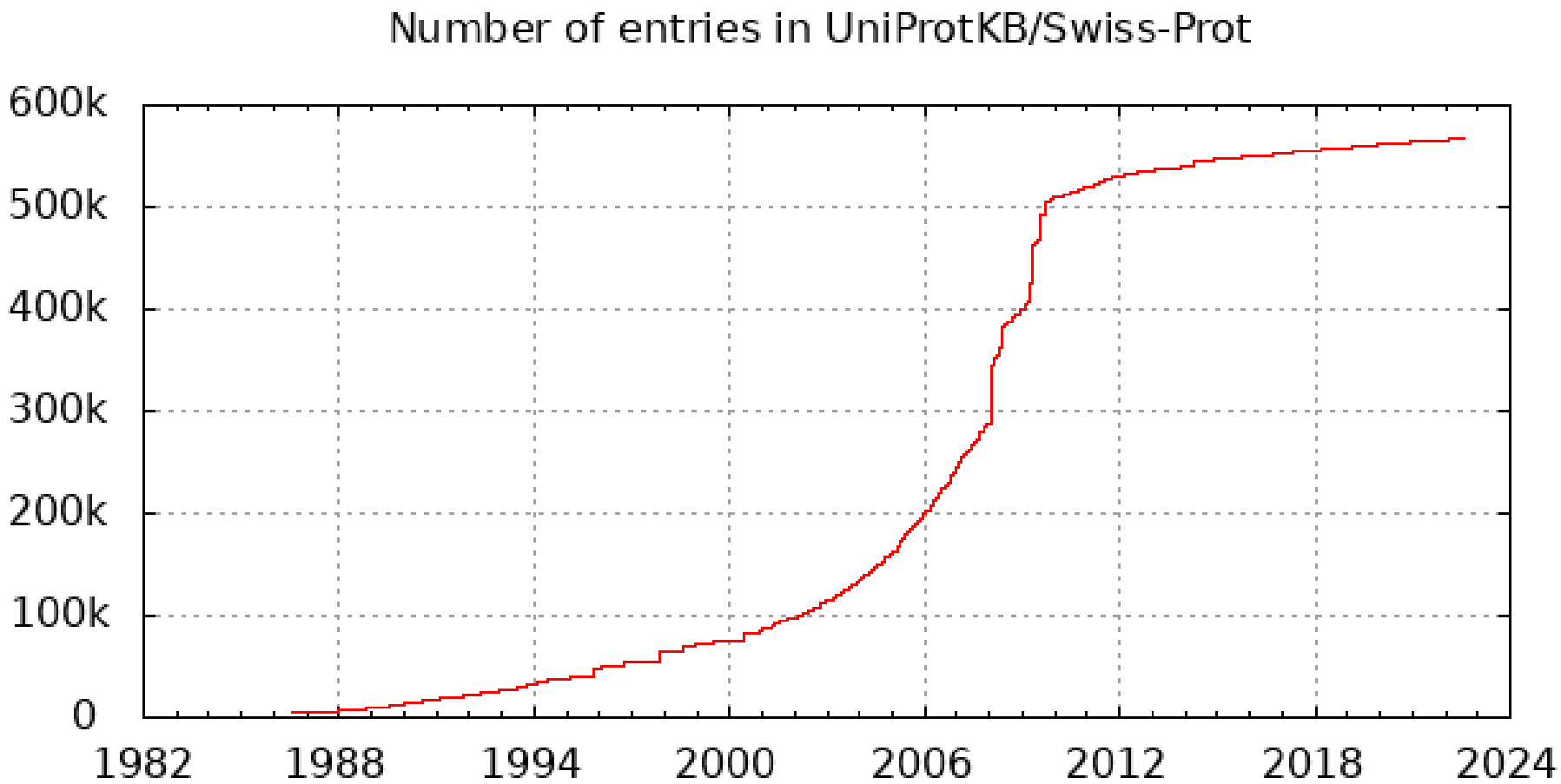
UniProt / SwissProt (manual annotation)

http://www.uniprot.org/



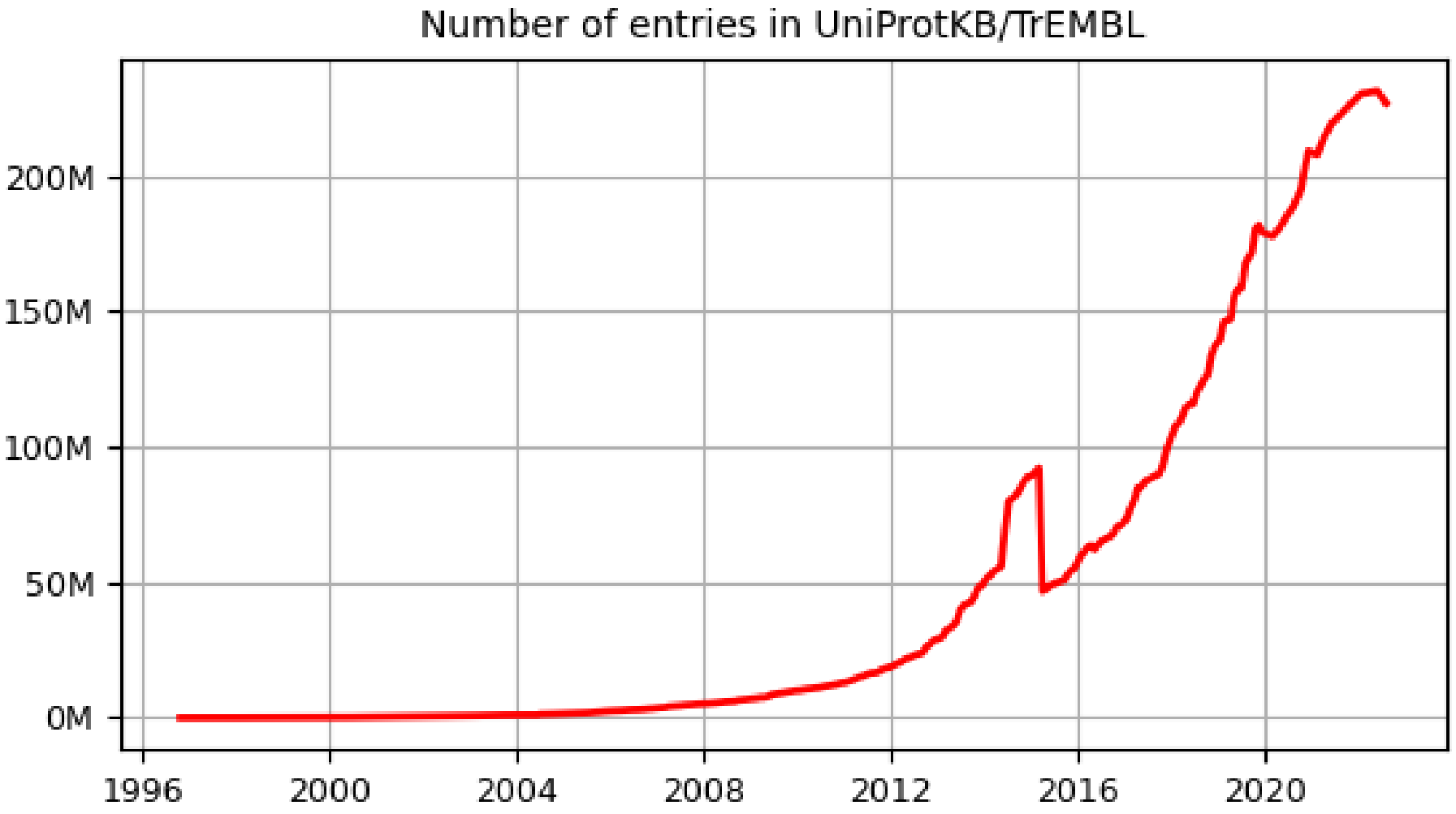
UniProt / SwissProt (manual annotation)

http://www.uniprot.org/



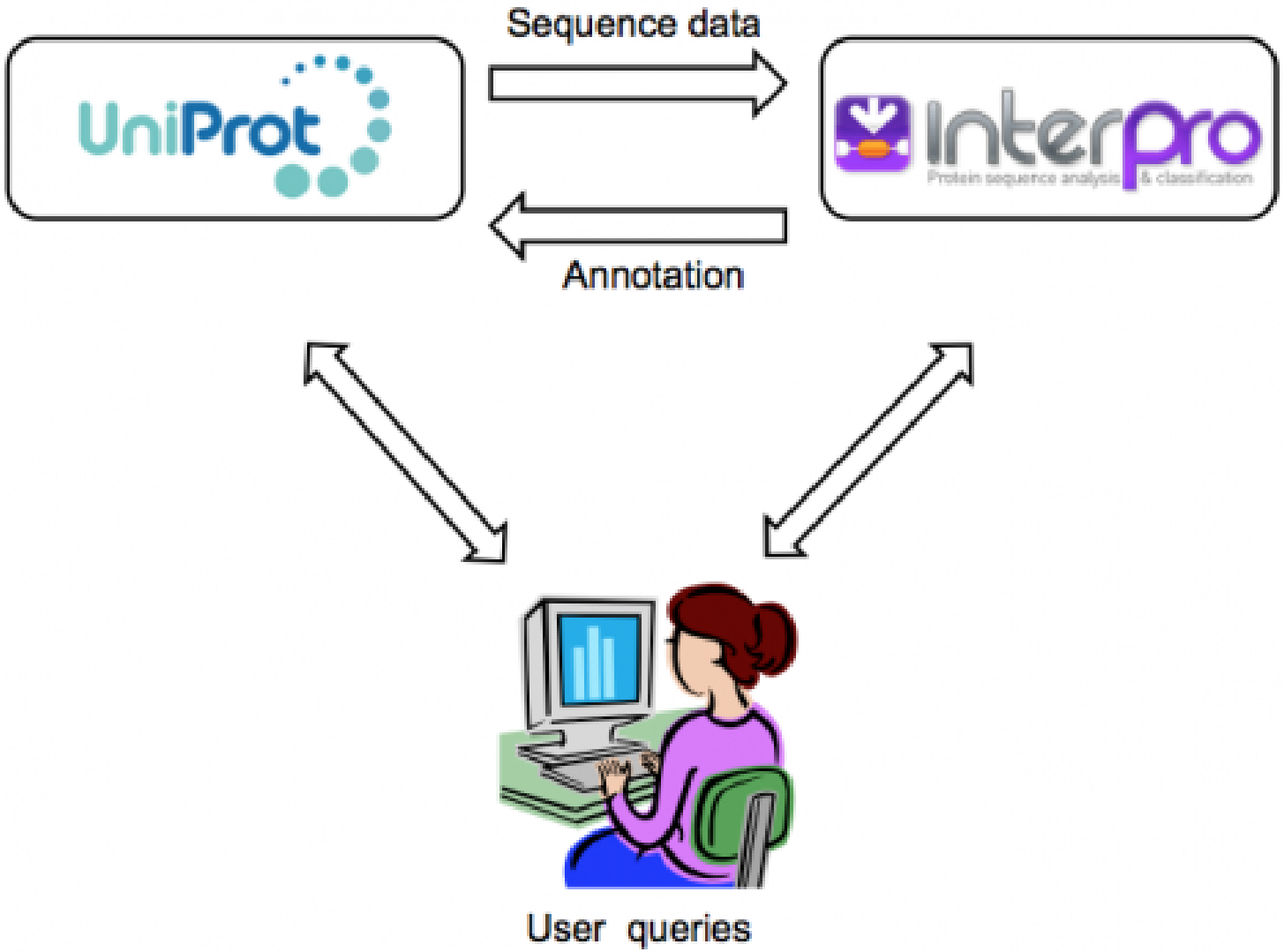
UniProt / TrEMBL (automatic annotation)

http://www.uniprot.org/



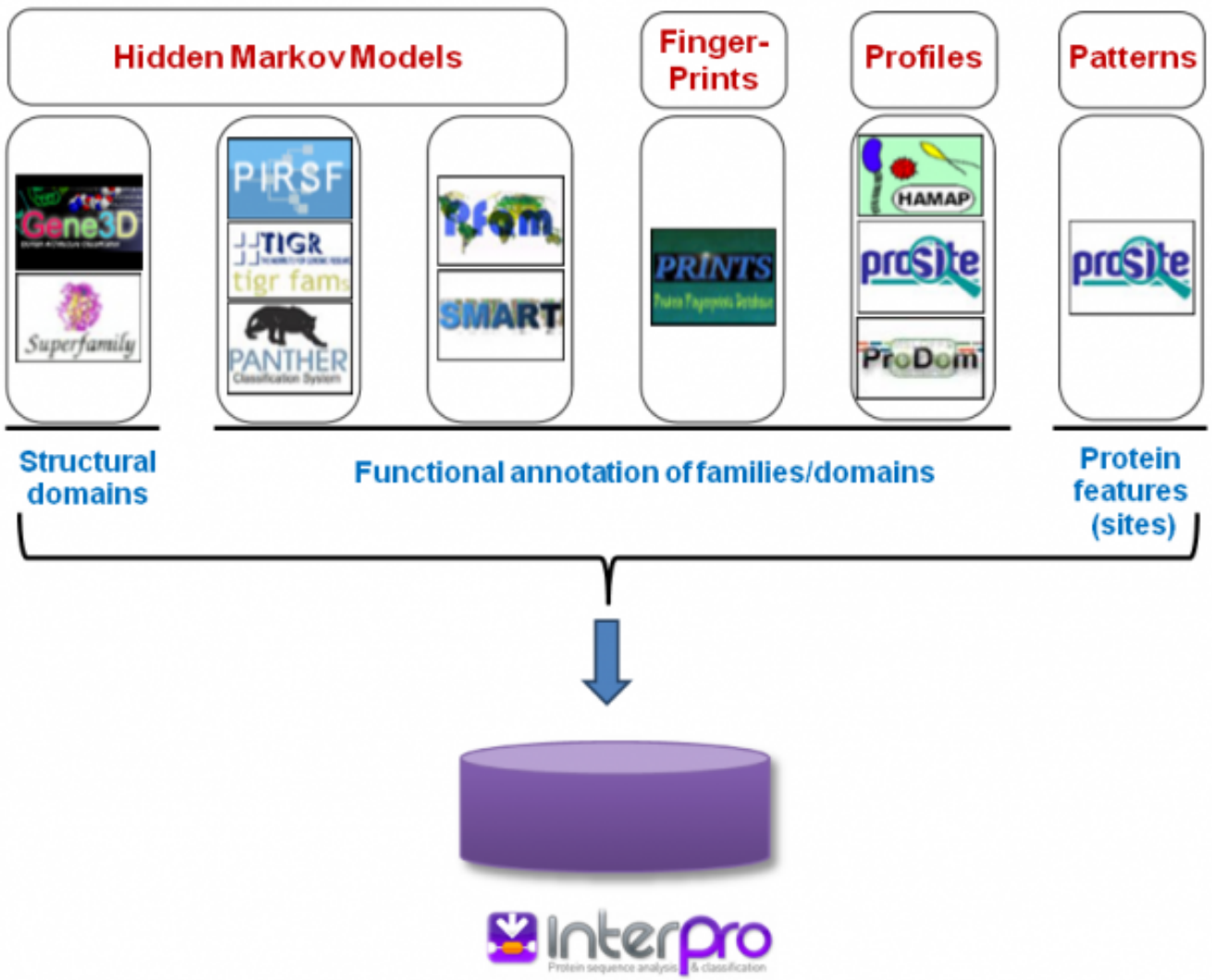
UniProt / TrEMBL (automatic annotation)

Done with InterPro: `http://www.ebi.ac.uk/interpro/`



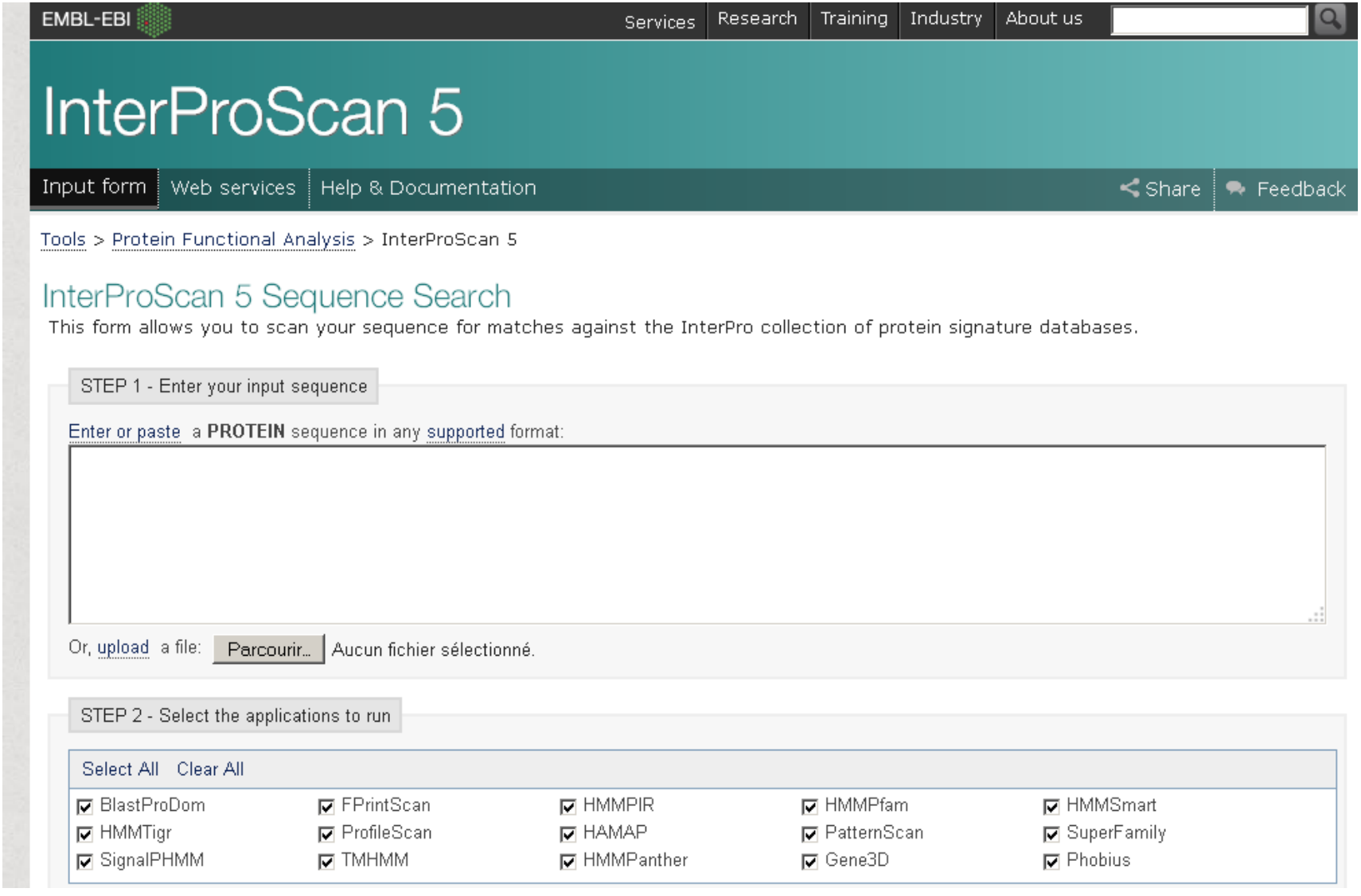
InterPro

http://www.ebi.ac.uk/interpro/



InterProScan

http://www.ebi.ac.uk/Tools/pfa/iprscan5/



Introduction alignement de séquences

Why Do We Want To Compare Sequences

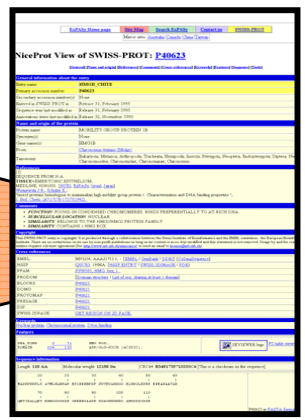
```
wheat   --DPNKPKRAMTSFVFFMSEFRSEFKQKHSKLKSIVEMVKAAGER
          | | ||||||||| || | |||    ||| |  |||| ||||
?????   KKDSNAPKRAMTSFMFFSSDFRS----KHSDL-SIVEMSKAAGAA
```

EXTRAPOLATE

Homology?

SwissProt

??????



Cédric Notredame (05/10/2013)

Introduction alignement de séquences

Why Does It Make Sense To Align Sequences ?

-Evolution is our Real Tool.

-Nature is LAZY and Keeps re-using Stuff.

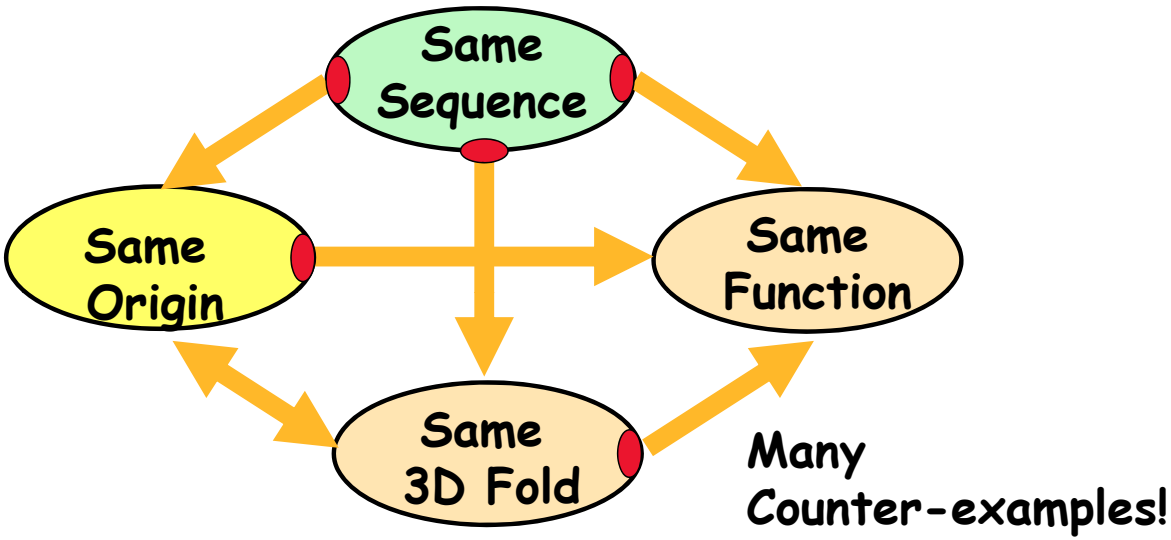
-Evolution is mostly DIVERGEANT

Same Sequence ⇔ Same Ancestor

Cédric Notredame (05/10/2013)

Introduction alignement de séquences

Why Does It Make Sense To Align Sequences ?



Cédric Notredame (05/10/2013)

Définitions

- *Identité*: Proportion de paires de résidus identiques entre 2 séquences. Dépend de l'alignement. Unité: %id
- *Similitude*: Proportion de paires de résidus similaires entre 2 séquences. Une matrice de substitution permet de décrire qui est similaire à qui (score > 0). Unité: %similarity
- *Homologie*: Deux séquences *similaires* peuvent être homologues si elles ont un ancêtre commun. Deux séquences avec peu de similarité peuvent aussi être homologues. Unité: Oui ou Non !

IL N'Y A PAS DE POURCENTAGE D'HOMOLOGIE : les séquences sont homologues ou elles ne le sont pas.

- Des séquences homologues ont souvent mais pas toujours la même fonction. . .
- . . . Elles ne sont pas forcément non plus très similaires : la structure est conservée plus que la séquence

Difficulté pour détecter l'homologie

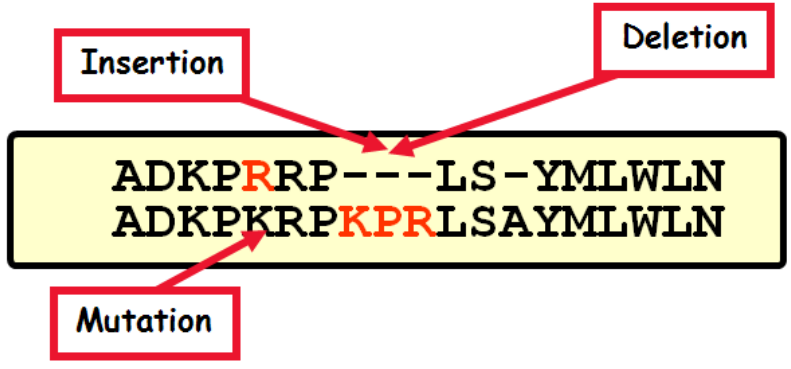
```
(a)
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
            G+ +VK+HGKKV  A+++++AH+D++ +++++LS+LH  KL
HBB_HUMAN   GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL

(b)
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL
            ++ ++++H+ KV   + +A  ++           +L+ L+++H+ K
LGB2_LUPLU  NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG

(c)
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSD----LHAHKL
            GS+ + G +   +D L  ++ H+ D+  A +AL D    ++AH+
F11G11.2    GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEFPQFKAHQE
```

(a) ok
(b) protéine homologue
(c) protéine non-homologue, mais même identité de séquence que (b)

Principes pour l'alignement des séquences

- Trouver des évidences que deux séquences ont divergées à partir d'un ancêtre commun => séquences homologues.
- Divergence: Processus de mutation et sélection

- Trois types de mutation:
  1. substitution
  2. insertion
  3. délétion

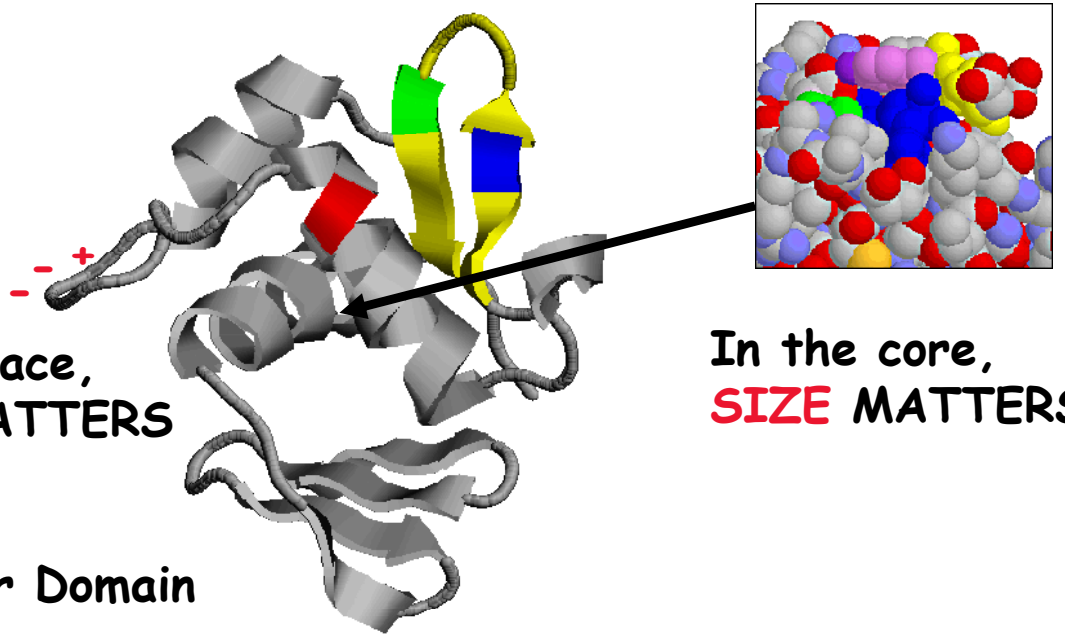


- Insertion/Délétion => *gaps* / trous
- Sélection naturelle favorise certaines mutations
- Score d'alignement = somme de termes pour chaque pair de résidus alignés + somme de termes pour chaque gap
- => Score additive => mutations à différents sites sont considérés comme indépendant

Mutations dépendent de la structure 3D

How Do Sequences Evolve ?

In a structure, each Amino Acid plays a Special Role



On the surface, CHARGE MATTERS

In the core, SIZE MATTERS

OmpR, Cter Domain

Cédric Notredame (05/10/2013)

Mutations dépendent de la structure 3D

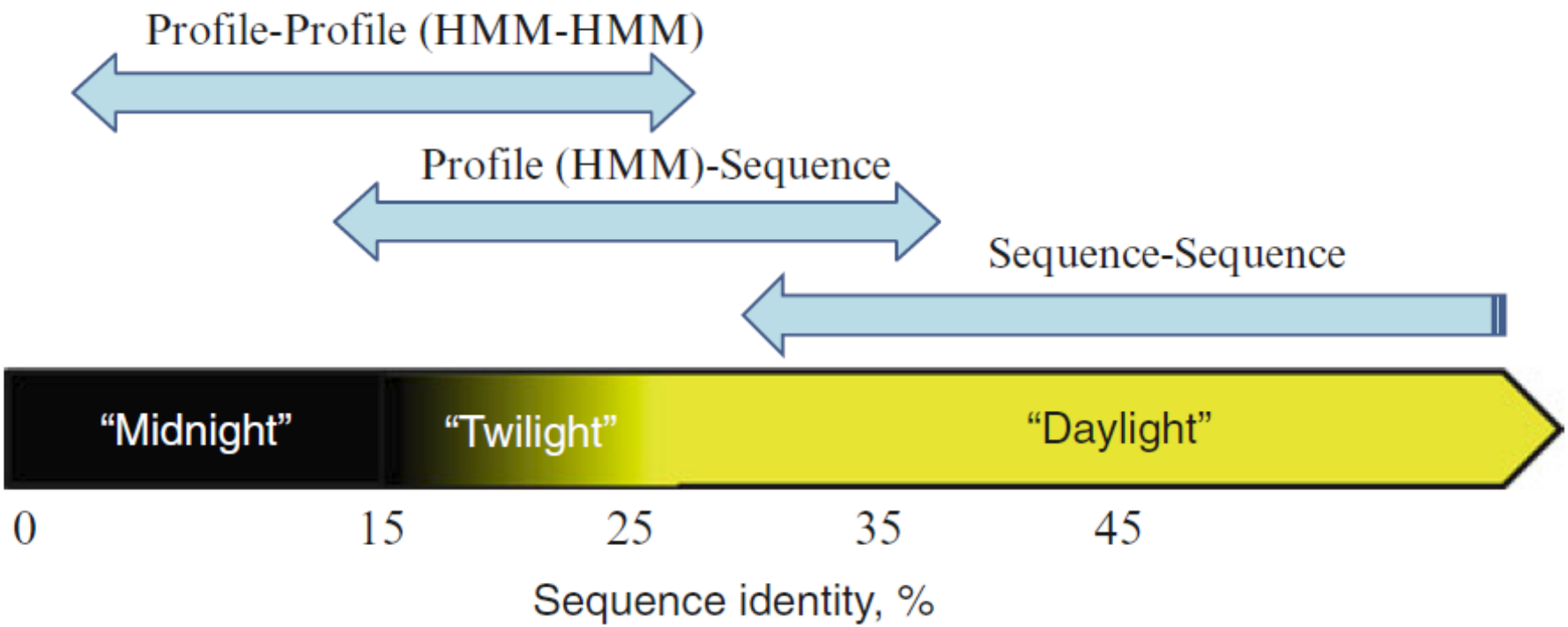
Le repliement d'une chaîne d'a.a. donne à chaque a.a. un environnement chimique qui dépend de la structure 3D du repliement.
Pour les protéines solubles:

- Surface => interface avec l'eau => a.a. polaires ou chargés
- Coeur => a.a. hydrophobes
- Sites actives ou de liaison => plus sensibles à la mutation

Betts and Russel, Bioinformatics for Geneticists, chapter 14

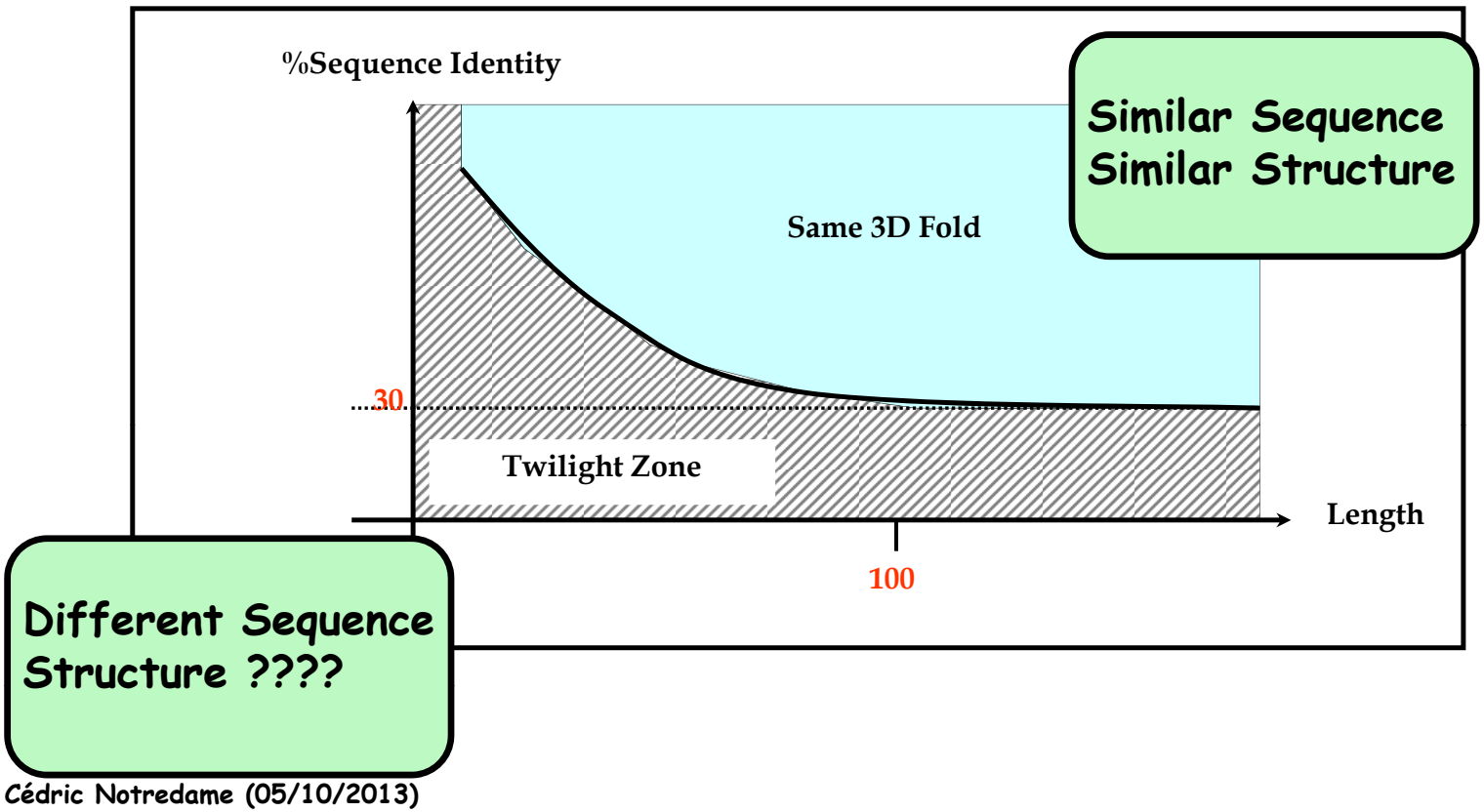
Alignement en pratique

Identité de séquence



Venclovas, "Homology Modeling", ch. 3, Methods in Mol. Biol.(2012)

How Can We Compare Sequences ?
The Twilight Zone



http://www.tcoffee.org/Courses/Exercises/upf_ug_2013/lectures/4.1.pairwise_intro.ppt

Alignement d'une paire de séquences

- Trois ingrédients:
  1. séquences d'acides aminés de deux protéines
  2. matrice avec des scores de substitution des résidus
  3. algorithme d'alignement
- Applicable à la "daylight" zone.
- Programmes: BLAST, FASTA
- La base des autres méthodes d'alignement (séquence-profil, profil-profil).
- Plus trop utilisé aujourd'hui dans la modélisation par homologie vu la supériorité des autres méthodes d'alignement.

Matrice de substitution dans BLAST

Choix de la matrice de substitution:

- Balance entre sensibilité et sélectivité
- *Sensibilité*: identifie des homologues lointaines, mais augmente les faux-positives
- *Sélectivité*: réduit les faux-positives, mais augmente le risque de rater des vrais homologues
- matrices BLOSUM: grand index (BLOSUM 80) = sélective, petit index (BLOSUM 45) = sensible

Variantes de BLAST

- CS-BLAST (context-specific)
  - score de substitution dépend des résidus voisins
  - prometteur pour les séquences "singleton" (séquences sans homologue détectable), car méthodes séquence-profil ou profil-profil ne marchent pas ici.
- PSI-BLAST (Position-specific iterated)
- PHI-BLAST (Pattern Hit Initiated BLAST): "performs the search but limits alignments to those that match a pattern in the query."
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST): "constructs a PSSM using the results of a Conserved Domain Database search and searches a sequence database."

PSI-BLAST (Position-specific iterated)

1. alignement multiple des meilleurs résultats d'une recherche initiale avec BLAST
2. à partir de cet alignement multiple: construction d'une matrice de score de substitution qui dépend de la position (position-specific scoring matrix (PSSM)).
3. nouvelle recherche avec BLAST et la matrice PSSM
4. répétition de l'étape 1) à 3) pour inclure des séquences de plus en plus éloigné.

Alignement séquence - profil ou HMM

- Informations extraites d'un alignment multiple et converties dans un modèle statistique compréhensive du groupe de séquences alignées.
  1. zones conservés ou variables
  2. zones avec insertions ou délétions
- Applicable à la "twilight" (15%-30% identité de séquence) et même "midnight" (<15%) zone.
- Programmes: PSI-BLAST, CSI-BLAST, HMMER
- Non traitée: effets qui dépendent de plusieurs positions (corrélations d'ordre supérieur), car chaque position dans la séquence est traitée indépendamment.

HMMER

- HMMs: Hiden Markov Models
- Comme les profils de séquence, mais le choix des scores est guidé par une théorie probabiliste.
- En plus HMMs contiennent des probabilités pour les insertions et délétions à chaque position du profil.
- Exemple: Le noyau structural d'une protéine est plus affecté par des insertions ou délétions que une boucle à la surface.

Alignement profil - profil ou HMM-HMM

- Comparaison de deux profils ou deux HMMs
- Question différente:
  - *Avant*: Est-ce que la séquence appartient à une famille et si oui, laquelle?
  - *Ici*: Est-ce que deux familles ont un lien évolutif?
- Permet de détecter des homologies malgré un faible taux d'identité de séquence ("midnight" zone)
- Plus juste en général que les méthodes d'alignement séquence-profil
- Programmes: HHsearch, PRC, PROCAIN

Méthodes pour détecter la homologie

Method	Type	Address
BLAST	Sequence–Sequence	http://blast.ncbi.nlm.nih.gov/
FASTA/Ssearch	Sequence–Sequence	http://fasta.bioch.virginia.edu/ http://www.ebi.ac.uk/Tools/sss/fasta/
CS-BLAST	Sequence (profile)–Sequence	http://toolkit.lmb.uni-muenchen.de/cs_blast/
PSI-BLAST	Profile–Sequence	http://blast.ncbi.nlm.nih.gov/
CSI-BLAST	Profile–Sequence	http://toolkit.lmb.uni-muenchen.de/cs_blast/
HMMER	HMM–Sequence	http://hmmer.org/
SAM	HMM–Sequence	http://compbio.soe.ucsc.edu/HMM-apps/
COMPASS	Profile–Profile	http://prodata.swmed.edu/compass/
PROCAIN	Profile–Profile + additional sequence features + SS[a]	http://prodata.swmed.edu/procain/
COMA	Profile–Profile	http://www.ibt.lt/bioinformatics/coma/
HHsearch	HMM–HMM + SS[a]	http://toolkit.lmb.uni-muenchen.de/hhpred/
PRC	HMM–HMM	http://supfam.org/PRC http://www.ibi.vu.nl/programs/prcwww/

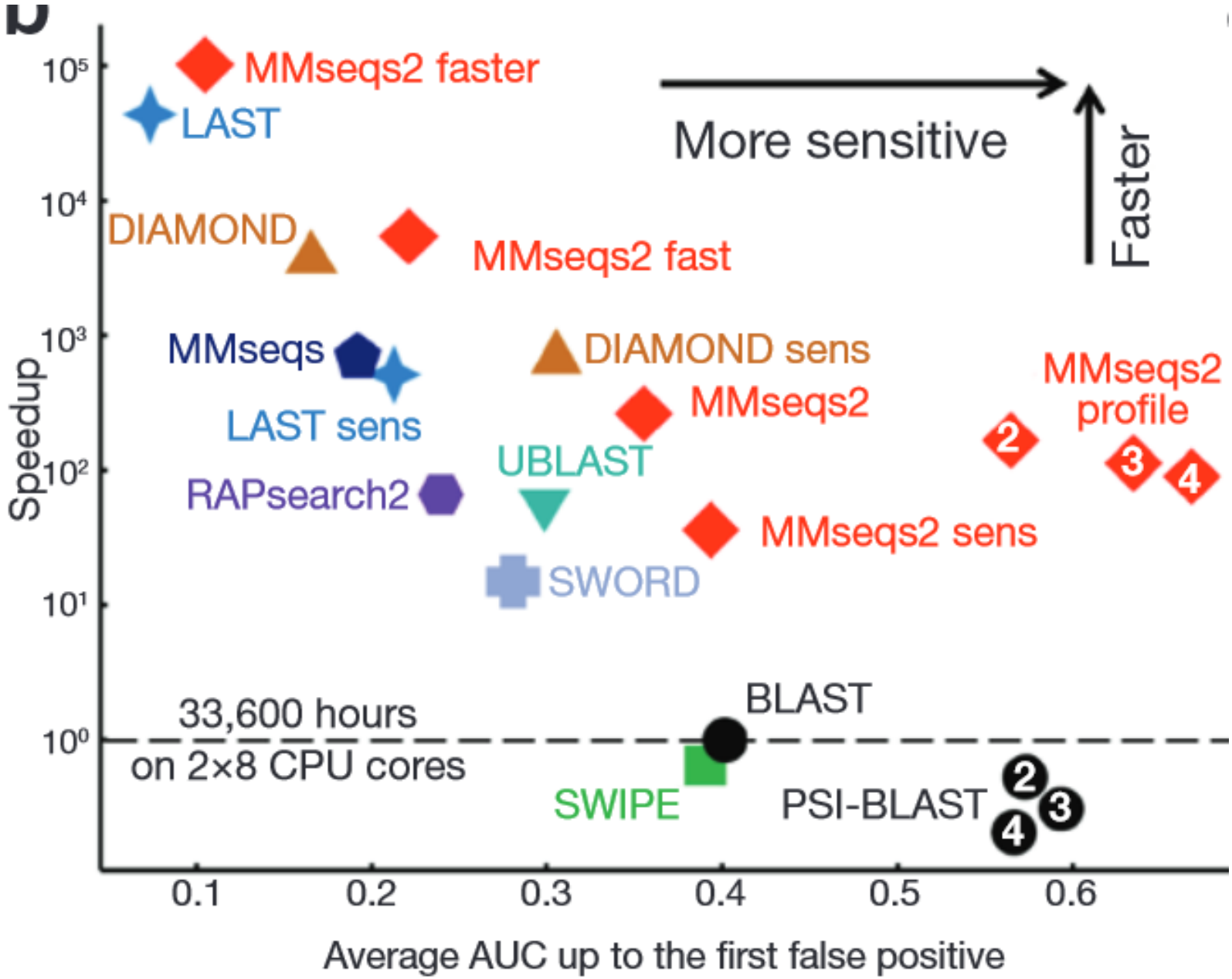
[a]Secondary structure

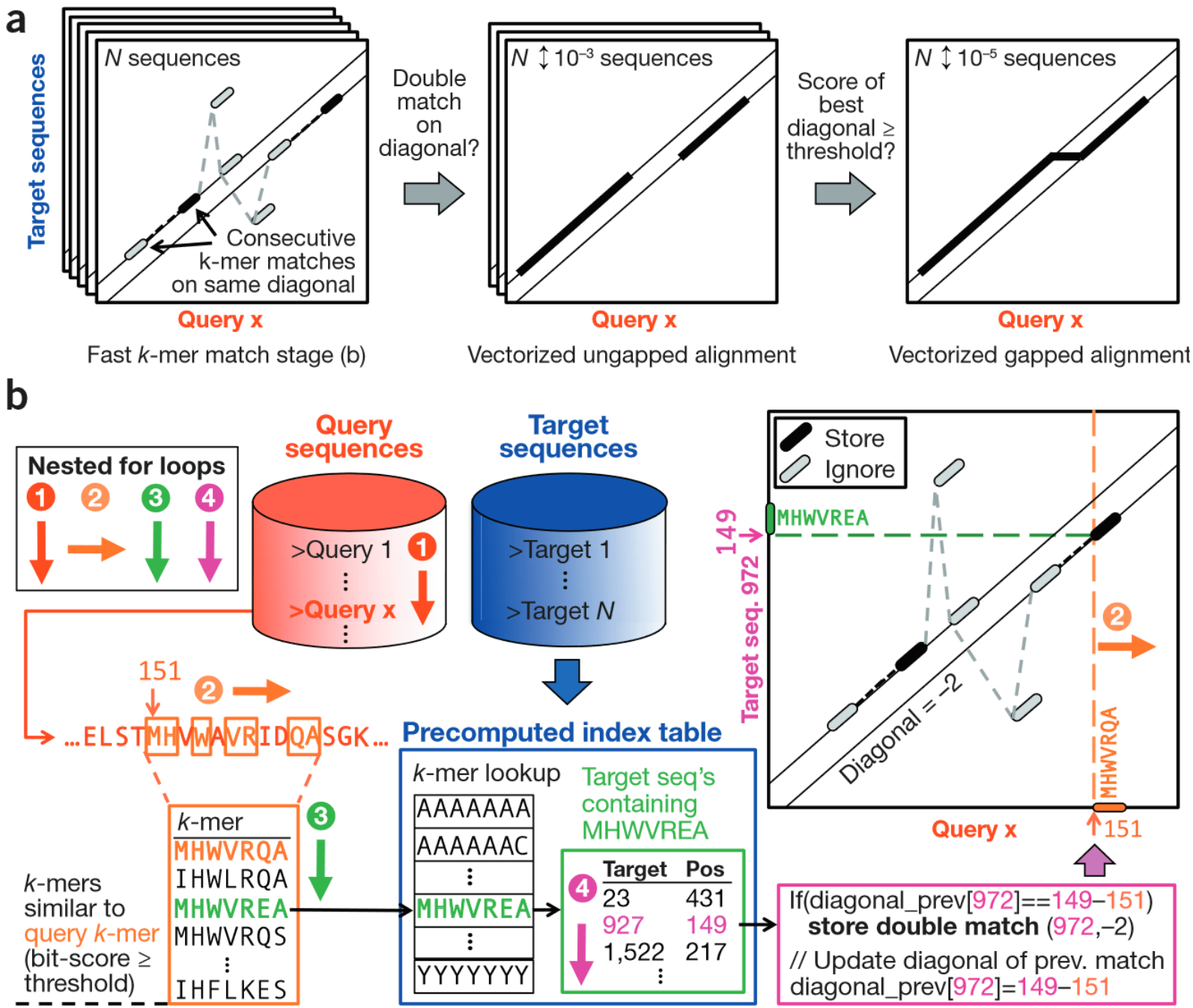
Venclovas, "Homology Modeling", ch. 3, Methods in Mol. Biol.(2012)

MMseqs2

- Coût du séquençage a diminué de 4 ordres de grandeurs depuis 2007
- Essort des projets de métagénomique
- Nouveau goulot d'étranglement: la traitement informatique de ces données en volume de Terabytes
- Une méthode récente beaucoup plus rapide que BLAST et PSI-BLAST et plus sensible en même temps: MMseqs2
- Des valeurs "E-value" plus justes, moins de faux positifs
- Utilisée dans colabfold (=AlphaFold plus rapide)

Martin Steinegger and Johannes Söding (Nov. 2017). en. In: *Nature Biotechnology* 35.11
Milot Mirdita, Martin Steinegger, and Johannes Söding (Aug. 2019). In: *Bioinformatics* 35.16
M Mirdita et al. (Sept. 2021). In: *Bioinformatics* 37.18
Milot Mirdita, Konstantin Schütze, et al. (June 2022). en. In: *Nature Methods* 19.6


MMseqs2 faster
LAST
More sensitive
Faster
DIAMOND
MMseqs2 fast
MMseqs
DIAMOND sens
MMseqs2
profile
LAST sens
MMseqs2
UBLAST
RAPsearch2
MMseqs2 sens
SWORD
BLAST
33,600 hours
on 2×8 CPU cores
SWIPE
PSI-BLAST
Speedup
Average AUC up to the first false positive


a
Target sequences
N sequences
Consecutive k-mer matches on same diagonal
Query x
Fast k-mer match stage (b)
Double match on diagonal?
N ↕ 10⁻³ sequences
Query x
Vectorized ungapped alignment
Score of best diagonal ≥ threshold?
N ↕ 10⁻⁵ sequences
Query x
Vectorized gapped alignment
b
Nested for loops
Query sequences
>Query 1
>Query x
Target sequences
>Target 1
>Target N
151
...ELSTMHWVREAVRIDQASGK...
k-mer
MHWVRQA
IHWLRQA
MHWVREA
MHWVRQS
IHFLKES
k-mers similar to query k-mer (bit-score ≥ threshold)
Precomputed index table
k-mer lookup
AAAAAAA
AAAAAAC
MHWVREA
YYYYYYY
Target seq's containing MHWVREA
Target Pos
23 431
927 149
1,522 217
Store
Ignore
149
MHWVREA
Target seq. 972
Diagonal = −2
MHWVRQA
Query x
151
If(diagonal_prev[972]==149–151)
store double match (972,–2)
// Update diagonal of prev. match
diagonal_prev[972]=149–151

Alignement multiple

2: Alignment correction

- Functional residues → conserved
- Use multiple sequence alignments
- Deletions → shift gaps

```
CPISRTAAS-FRCW
CPISRTG-SMFRCW
CPISRTA--TFRCW
CPISRTAASHFRCW
CPISRTGASIFRCW
CPISRTA---FRCW
```

Multipe sequence alignment

```
CPISRTGASIFRCW    CPISRTGASIFRCW
CPISRTA---FRCW    CPISRT---AFRCW
```

← Sequence with known structure

←Your sequence

Correct alignment

Hanka Venselaar, http://swift.cmbi.ru.nl/teach/B1SEM/B1S_homology_modelling.ppt

Introduction

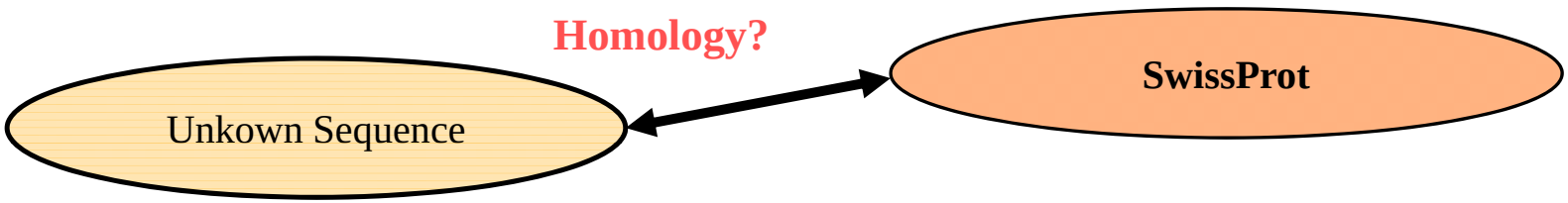
- Les méthodes d'alignement multiple (multiple sequence alignment (MSA)) ne sont pas fait pour détecter des séquences homologues
- Elles permettent d'aligner un jeu de séquences homologues identifié au préalable avec les méthodes d'alignement simple
- Grand nombre d'applications en biologie, dont:
  1. Reconstruction phylogénétique
  2. Construction des profiles (= matrices de substitution dépendantes de la position)
  3. Modélisation par homologie: Si la cible et le template sont dans le jeu de séquences à aligner, on peut obtenir leur alignement depuis l'alignement multiple
  4. Contraintes de distances par co-évolution: ex. AlphaFold

How Can I Use A Multiple Sequence Alignment?

```
chite     ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat     --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr     KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
unknown   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
                ***. ::: .: ..  .    :  . .     *  .  *: *

chite     AATAKQNYIRALQEYERNGG-
wheat     ANKLKGEYNKAIAAYNKGESA
trybr     AEKDKERYKREM---------
unknown   AKDDRIRYDNEMKSWEEQMAE
          *   : .* . :
```

Less Than 30 % id
BUT
Conserved where it MATTERS

Extrapolation Beyond The Twilight Zone



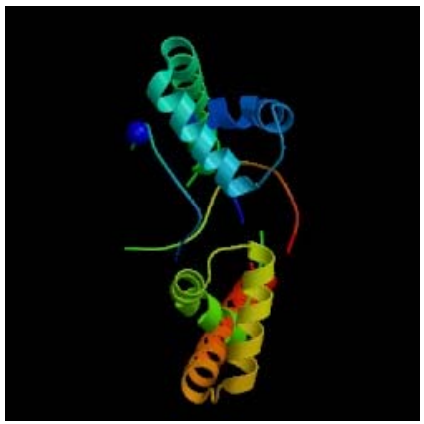
http://www.tcoffee.org/Courses/Exercises/upf_ug_2013/lectures/5.1.msa_introduction.ppt

How Can I Use A Multiple Sequence Alignment?

```
chite   ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKD
wheat   --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSE
trybr   KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGP
mouse   -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSP
             ***. ::: .: ..    :  . .     *  .  *: *

chite   AATAKQNYIRALQEYERNGG-
wheat   ANKLKGEYNKAIAAYNKGESA
trybr   AEKDKERYKREM---------
mouse   AKDDRIRYDNEMKSWEEQMAE
        *   : .* . :
```

- Extrapolation
- Motifs/Patterns
- Profiles
- Phylogeny
- Struc. Prediction

Column Constraint
⇔
Evolution Constraint
⇔
Structure Constraint



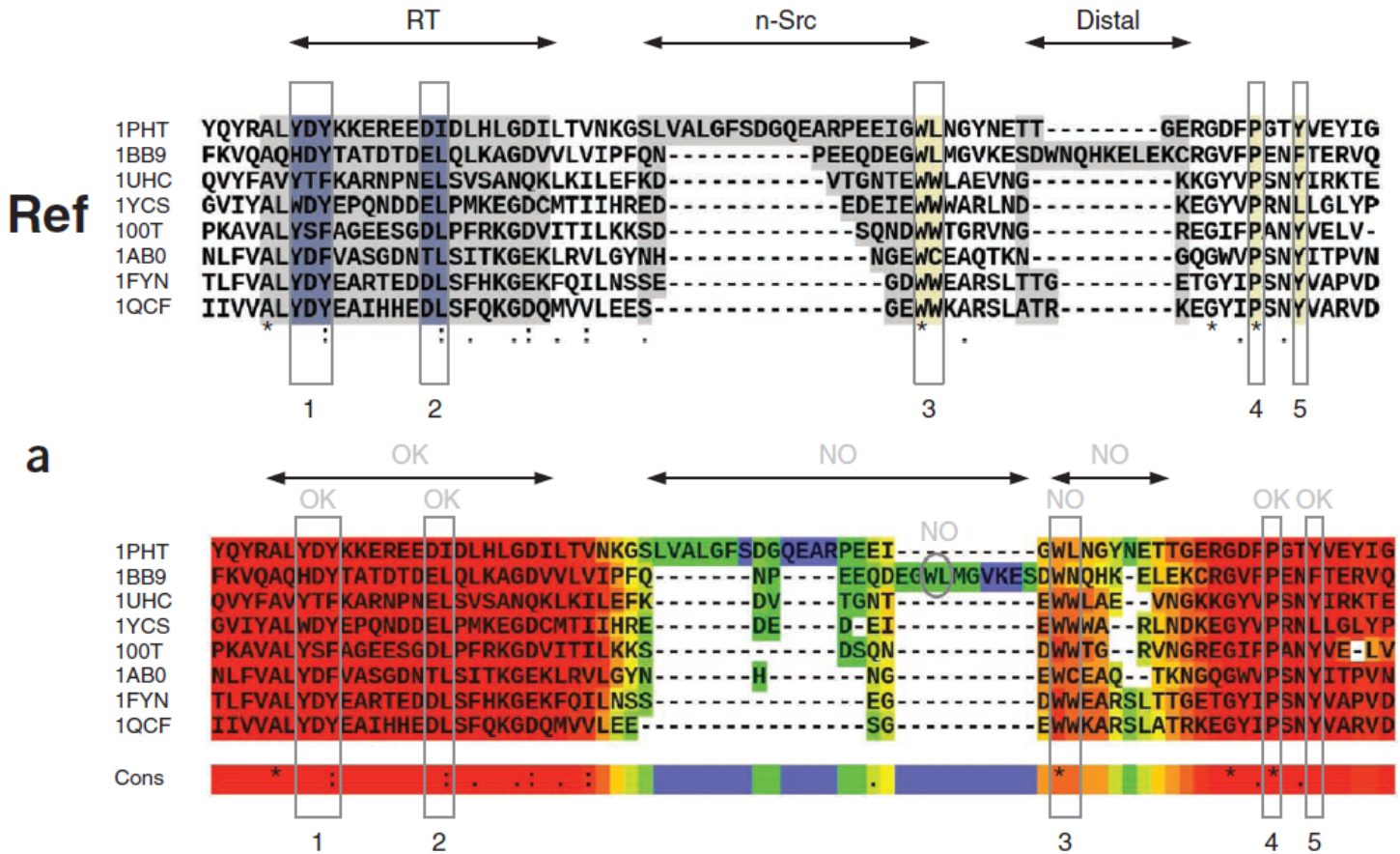
http://www.tcoffee.org/Courses/Exercises/upf_ug_2013/lectures/5.1.msa_introduction.ppt

Reading Your Alignment

* **A star indicates an entirely conserved column.**
: **A semi column indicates columns where all the residues have roughly the same size and the same hydropathy.**
. **A period indicates columns were the size OR the hydropathy has been preserved in the course of evolution.**

MSA - difficile

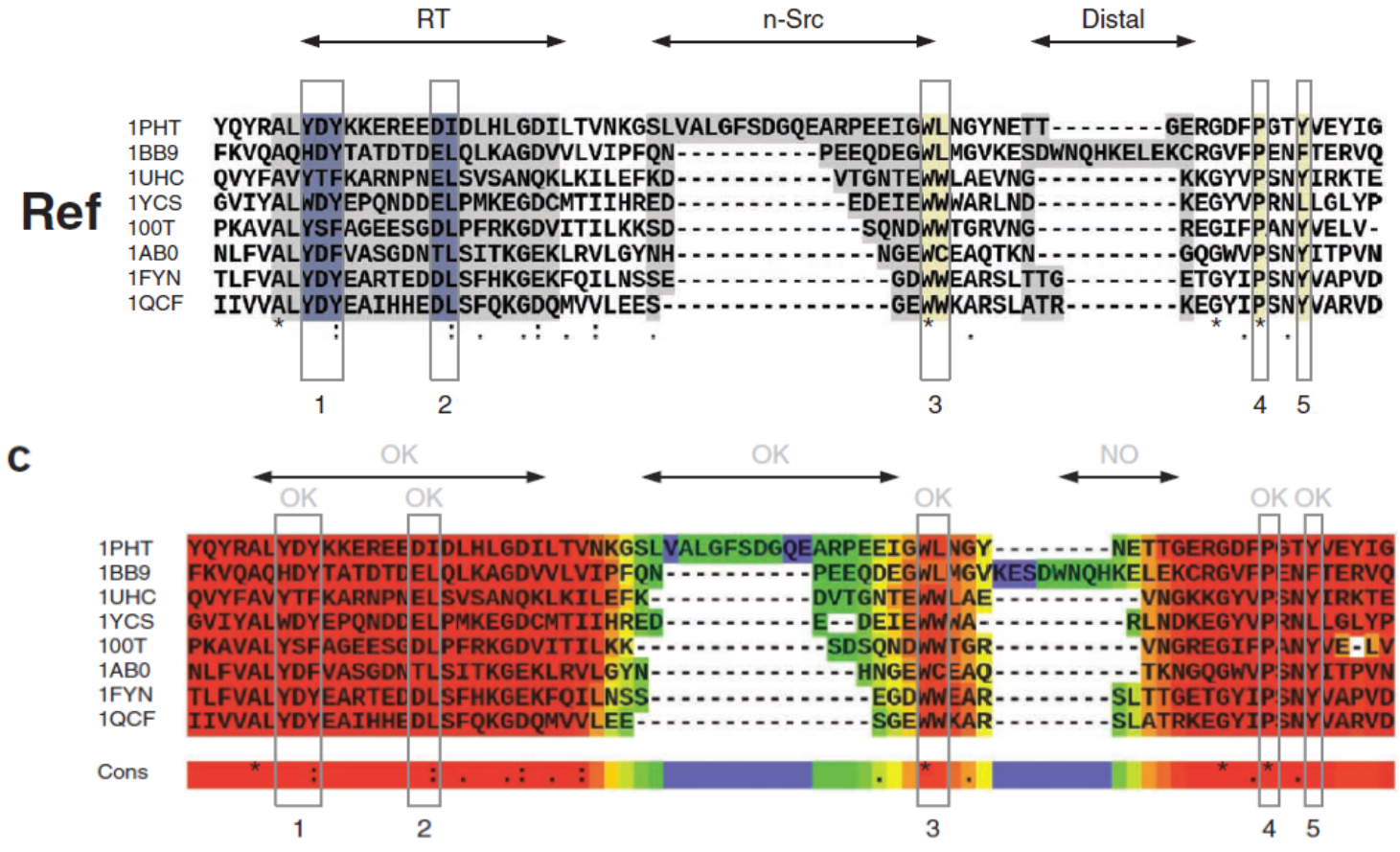
Exemple: domaines SH3, Ref: manuel, a: T-Coffee



Jean-Francois Taly et al. (Nov. 2011). en. In: *Nature Protocols* 6.11

MSA - mieux avec *homology extension*

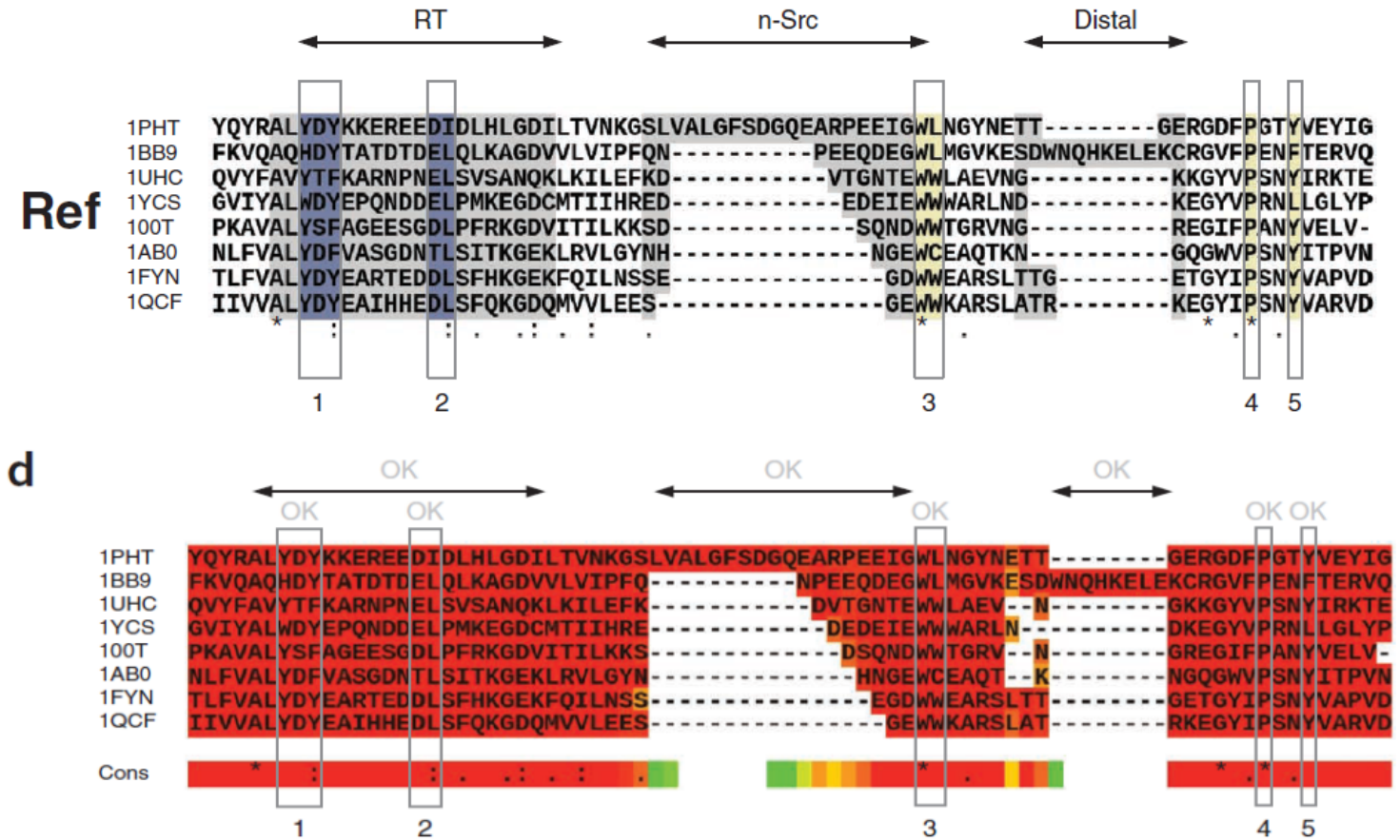
Exemple: domaines SH3, Ref: manuel, d: PSI-Coffee



Jean-Francois Taly et al. (Nov. 2011). en. In: *Nature Protocols* 6.11

MSA - plus simple avec structures 3D

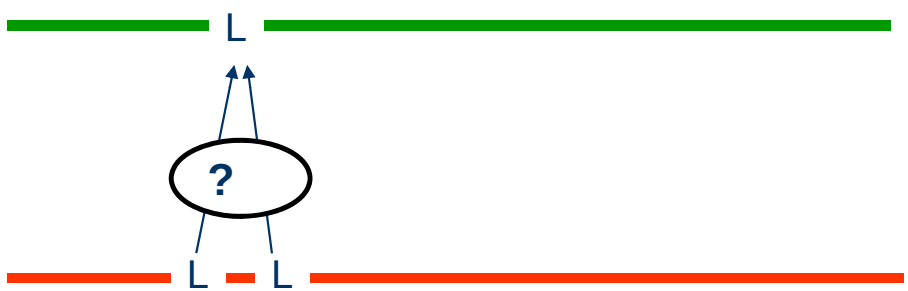
Exemple: domaines SH3, Ref: manuel, d: Expresso



Jean-Francois Taly et al. (Nov. 2011). en. In: *Nature Protocols* 6.11

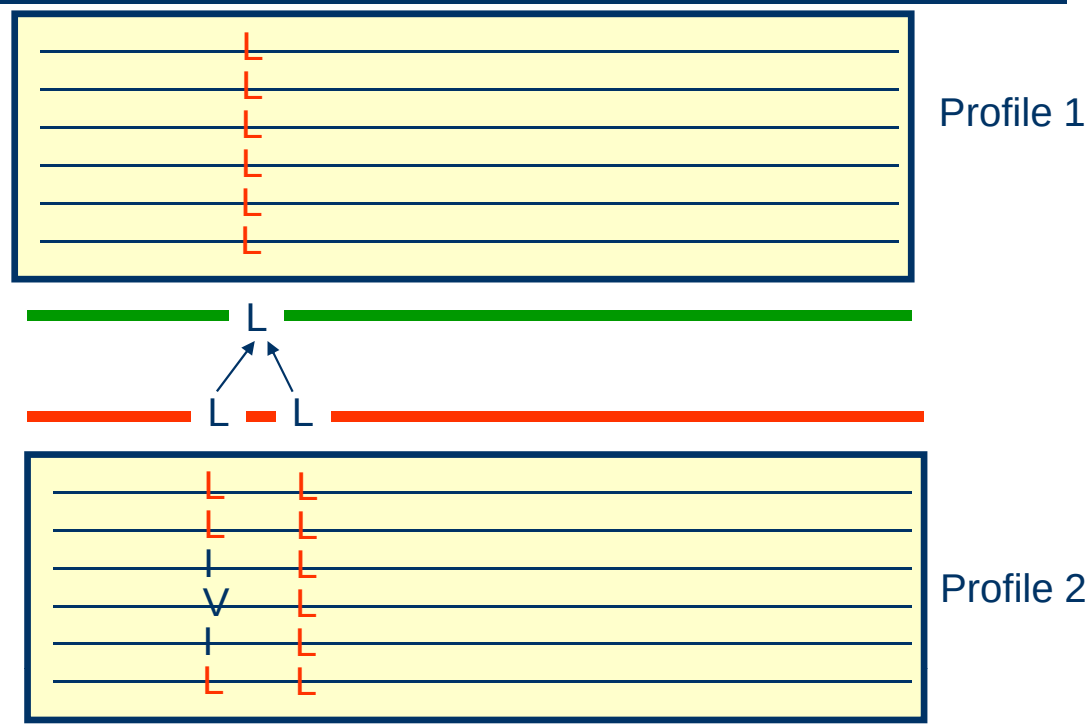
What is Homology Extension ?

-Simple scoring schemes result in alignment ambiguities



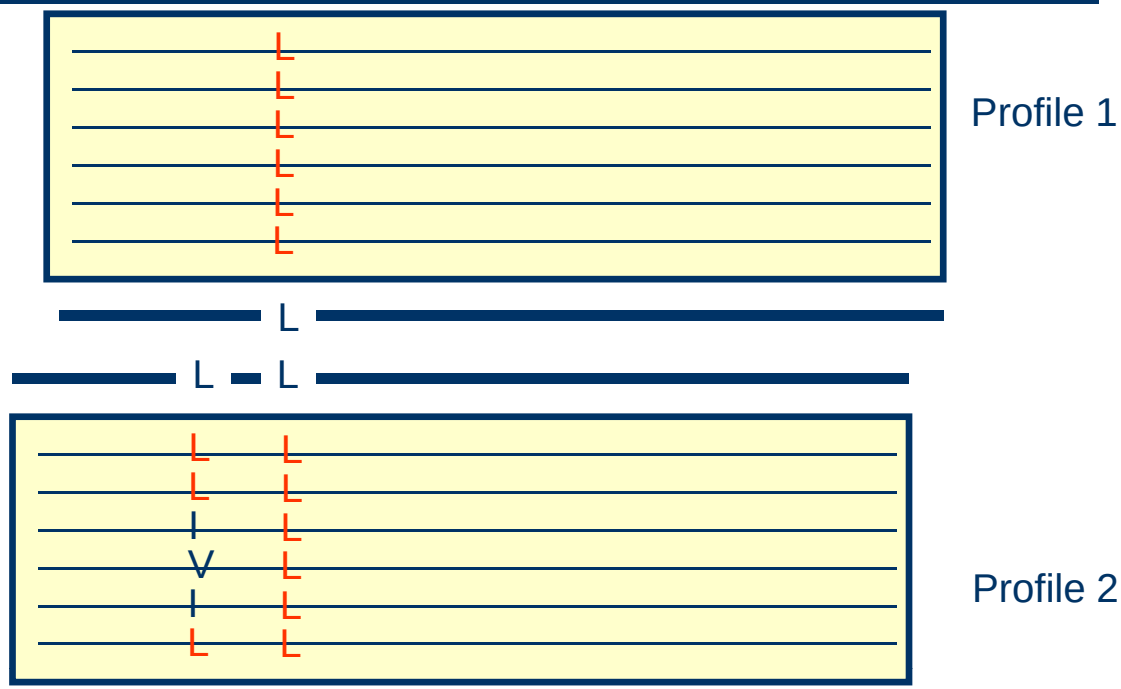
from Notredame, http://www.tcoffee.org/Presentations/T_Coffee.Workshop.CRG.ppt

What is Homology Extension ?



from Notredame, `http://www.tcoffee.org/Presentations/T_Coffee.Workshop.CRG.ppt`

What is Homology Extension ?



from Notredame, http://www.tcoffee.org/Presentations/T_Coffee.Workshop.CRG.ppt

Algorithmes / Programmes

1. *Alignement progressive*: ClustalW (ancien, mais toujours populaire)
2. *Affinement itérative*: MAFFT, MUSCLE
3. *Information de cohérence / consistency information*: T-coffee, ProbCons
   Plus juste mais aussi plus lente que l'algorithme (2)
4. *Combinaison de différentes méthodes*: M-coffee
5. *Avec alignement de structures 3D*: PROMALS3D, 3DCoffee/Expresso
6. *Edition manuelle*: JalView

What is the Best MSA method ?

- More than 50 MSA methods
- Some methods are fast and inacurate
  - Mafft, muscle, kalign
- Some methods are slow and accurate
  - T-Coffee, ProbCons
- Some Methods are slow and inacurate…
  - ClustalW

from Notredame, `http://www.tcoffee.org/Presentations/T_Coffee.Workshop.CRG.ppt`

Method	Method	Template	Score	Comment
ClustalW-2	Progressive	NO	**22.74**	
PRANK	Gap	NO	**26.18**	Science2008
MAFFT	Iterative	NO	**26.18**	
Muscle	Iterative	NO	**31.37**	
ProbCons	Consistency	NO	**40.80**	
ProbCons	MonoPhasic	NO	**37.53**	
T-Coffee	Consistency	NO	**42.30**	
M-Coffe4	Consistency	NO	**43.60**	
PSI-Coffee	**Consistency**	**Profile**	**53.71**	
PROMAL	**Consistency**	**Profile**	**55.08**	
PROMAL-3D	**Consistency**	**PDB**	**57.60**	
3D-Coffee	**Consistency**	**PDB**	**61.00**	**Expresso**

Score: fraction of correct columns when compared with a structure based reference (BB11 of BaliBase).

from Notredame, `http://www.tcoffee.org/Presentations/T_Coffee.Workshop.CRG.ppt`

Comparaisons plus récentes des méthodes MSA

Comparaisons très détaillées:

Mathilde Carpentier and Jacques Chomilier (Oct. 2019). In: *Bioinformatics* 35.20

Chapitre d'un livre récent sur MSA:

Tandy Warnow (2021). en. In: *Multiple Sequence Alignment: Methods and Protocols*. Methods in Molecular Biology

AlphaFold2 utilise FAMSA:

Sebastian Deorowicz, Agnieszka Debudaj-Grabysz, and Adam Gudyś (Sept. 2016). en. In: *Scientific Reports* 6.1

Vice-versa, AlphaFold2 aide pour améliorer les MSA:

Athanasios Baltzis et al. (Sept. 2022). In: *Bioinformatics*

3 Modélisation par homologie

- Modélisation / Prédiction d'une structure 3D de protéine
- Modélisation par homologie - Introduction
- Sélection du template
- Alignement cible - template
- Construction du modèle
- Évaluation du modèle

Modélisation / Prédiction d'une structure 3D de protéine

Modélisation par homologie - Introduction

Définition

- Génération d'un modèle 3D à partir d'un alignement de séquences de structure 3D connue
- Modéliser, c'est prédire !
- Noms en anglais: Homology modeling, Comparative modeling, Template-based modeling (TBM)
- template = support 3D ou patron
- Protéines / Domaines homologues:
  Protéines / Domaines avec lien(s) évolutif

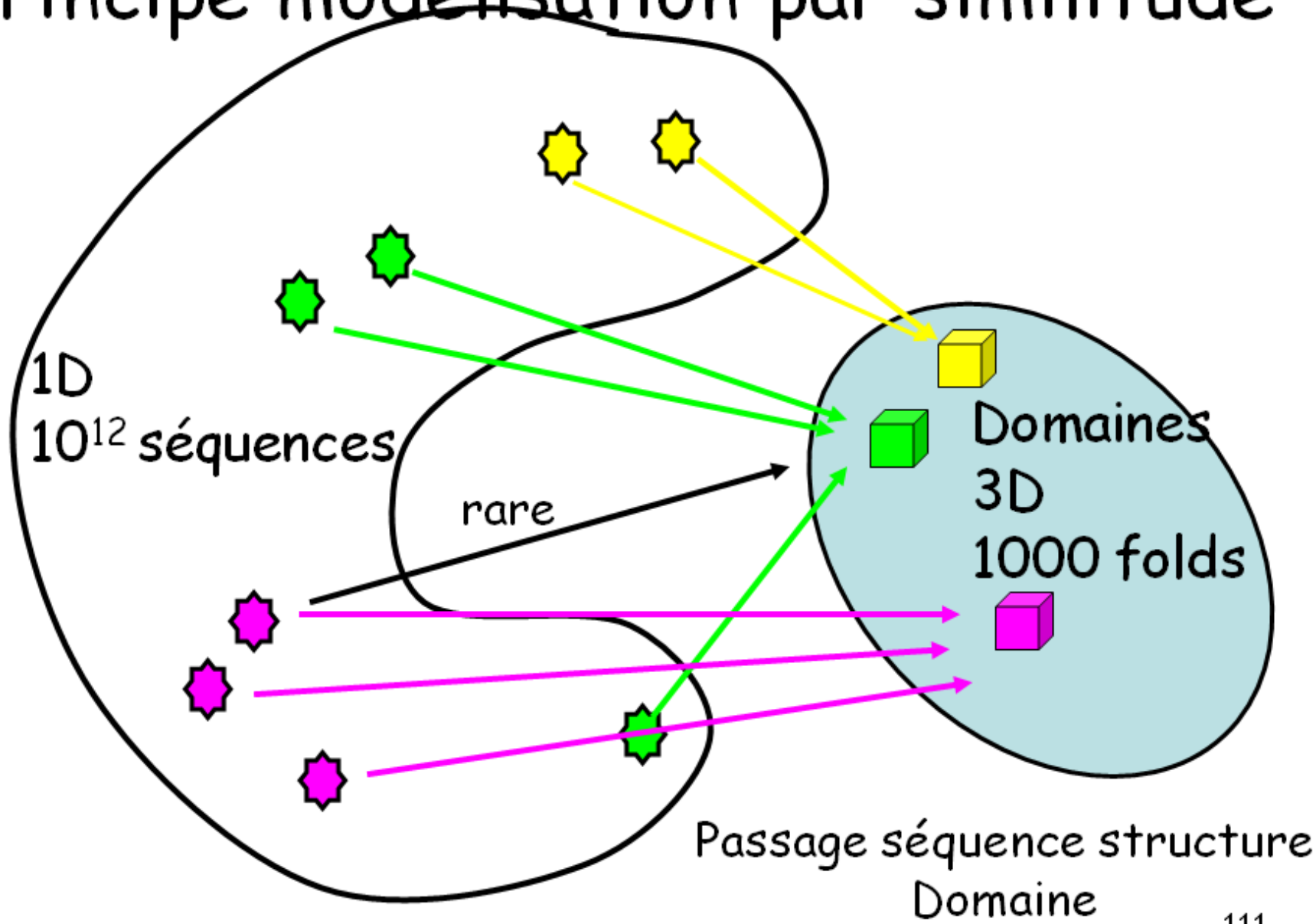
Méthode

Observation empirique:

- Domaines de protéines avec lien(s) évolutif ont la tendance d'avoir une structure 3D similaire.
- Structure 3D est mieux conservée que la séquence ou la fonction.
- Il y a des exceptions, mais cette règle reste vrai pour la majorité absolue des cas.

Parmi les méthodes de prédiction de structure, la modélisation par homologie est la plus précise, donc la plus utilisée.

Principe modélisation par similitude



Homology modeling in short…

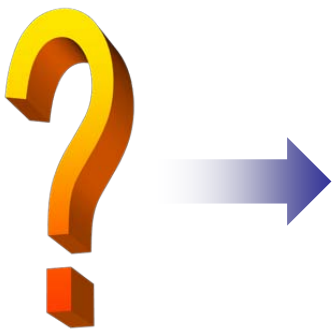
Prediction of structure based upon a highly similar structure

Hanka Venselaar, http://swift.cmbi.ru.nl/teach/B1SEM/B1S_homology_modelling.ppt

Homology modeling in short…

Prediction of structure based upon a highly similar structure

NSDSECPLSHDG



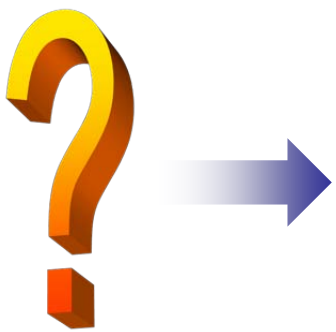
Unknown structure

Hanka Venselaar, http://swift.cmbi.ru.nl/teach/B1SEM/B1S_homology_modelling.ppt

Homology modeling in short…

Prediction of structure based upon a highly similar structure

NSDSECPLSHDG



Unknown structure

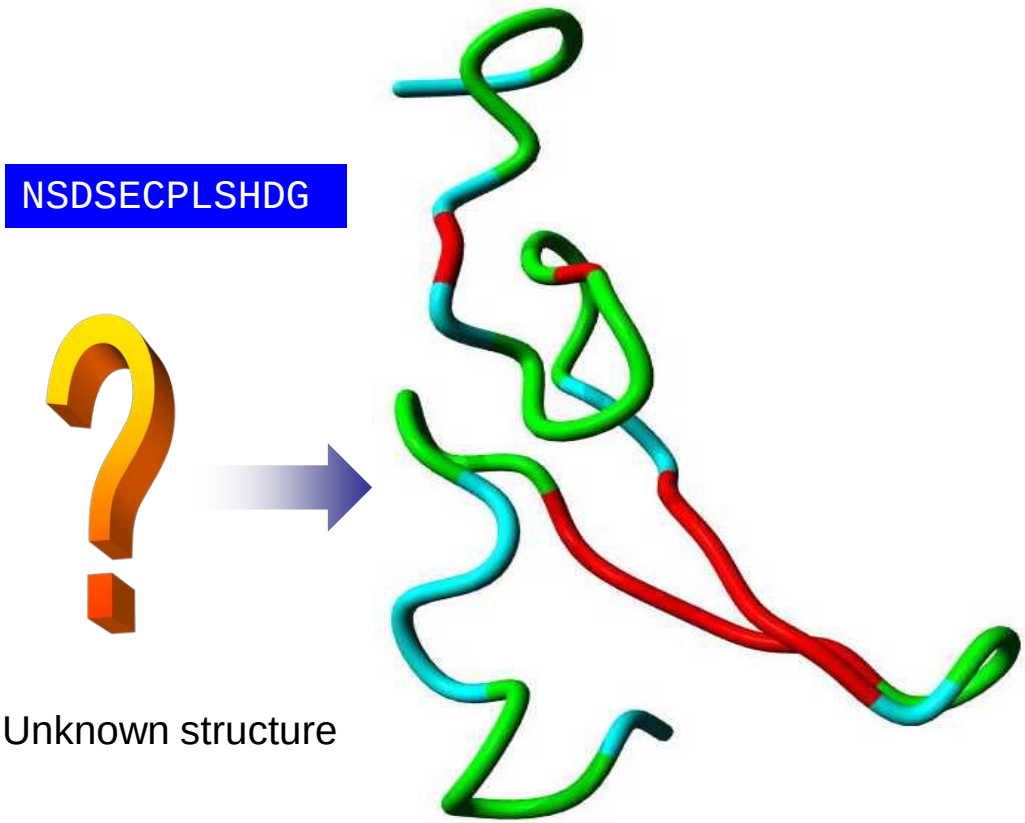
```
NSDSECPLSHDG
||   || | ||
NSYPGCPSSYDG
```

Alignment of model and template sequence

Hanka Venselaar, http://swift.cmbi.ru.nl/teach/B1SEM/B1S_homology_modelling.ppt

Homology modeling in short…

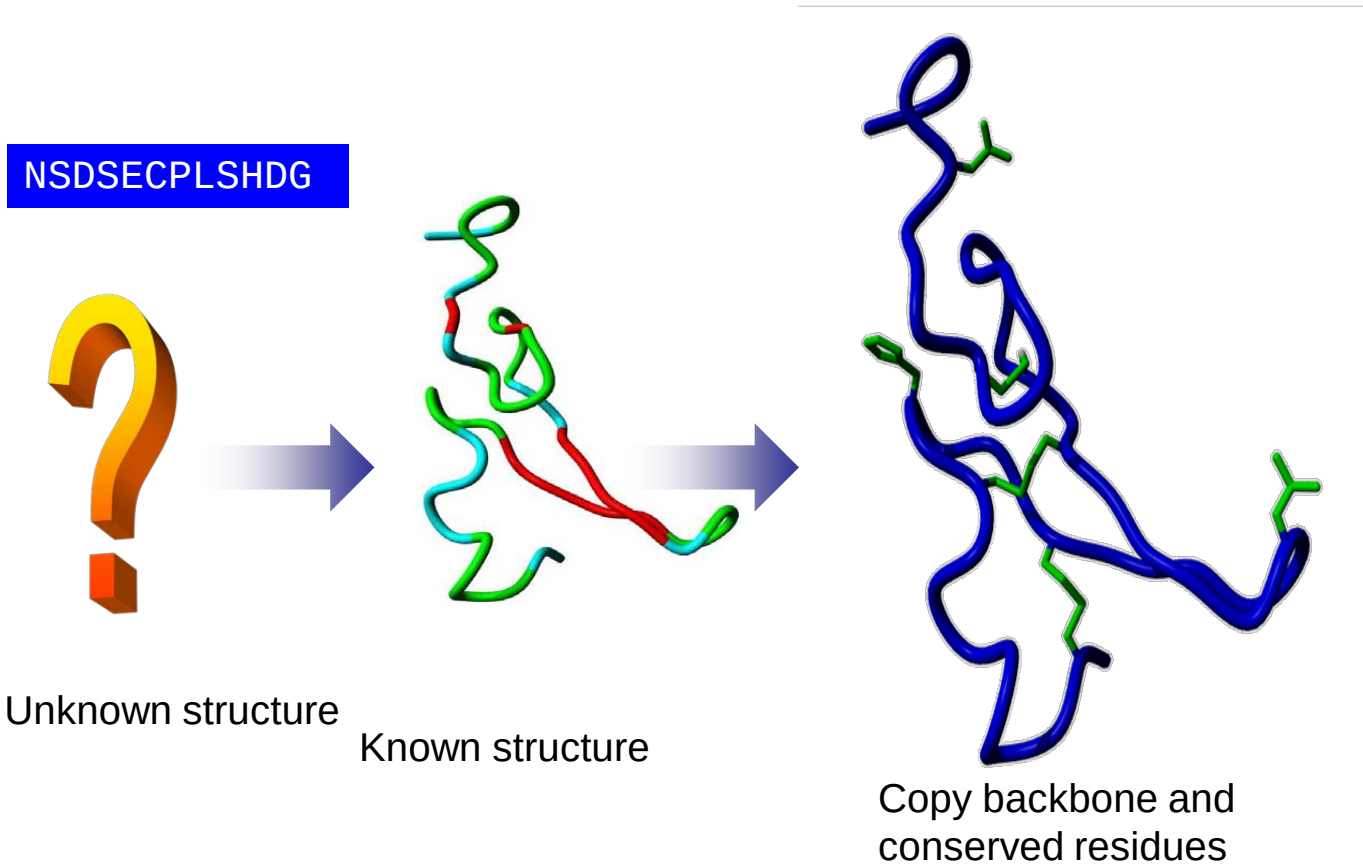
Prediction of structure based upon a highly similar structure



Hanka Venselaar, http://swift.cmbi.ru.nl/teach/B1SEM/B1S_homology_modelling.ppt

Homology modeling in short…

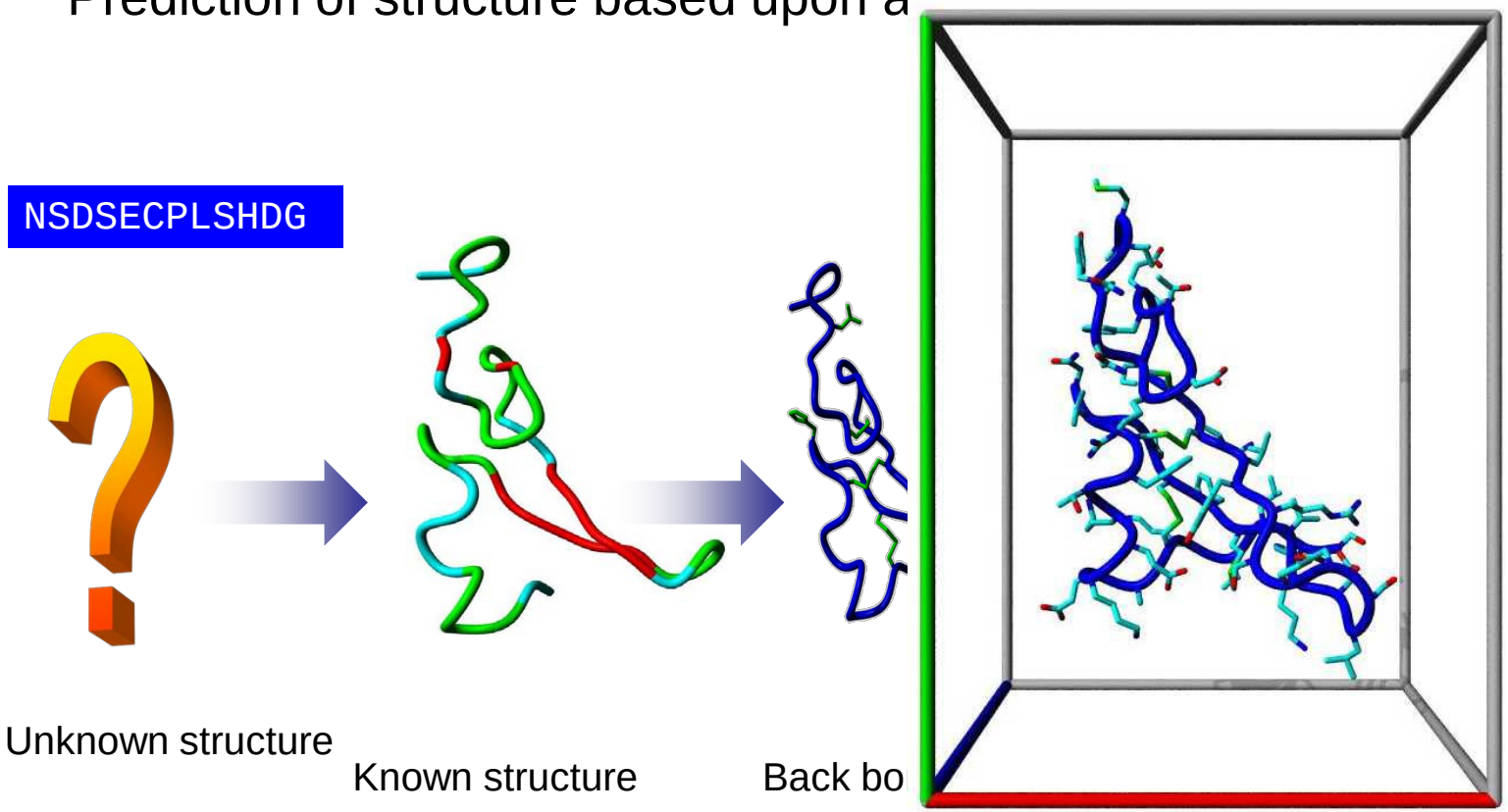
Prediction of structure based upon a highly similar structure



Hanka Venselaar, `http://swift.cmbi.ru.nl/teach/B1SEM/B1S_homology_modelling.ppt`

Homology modeling in short…

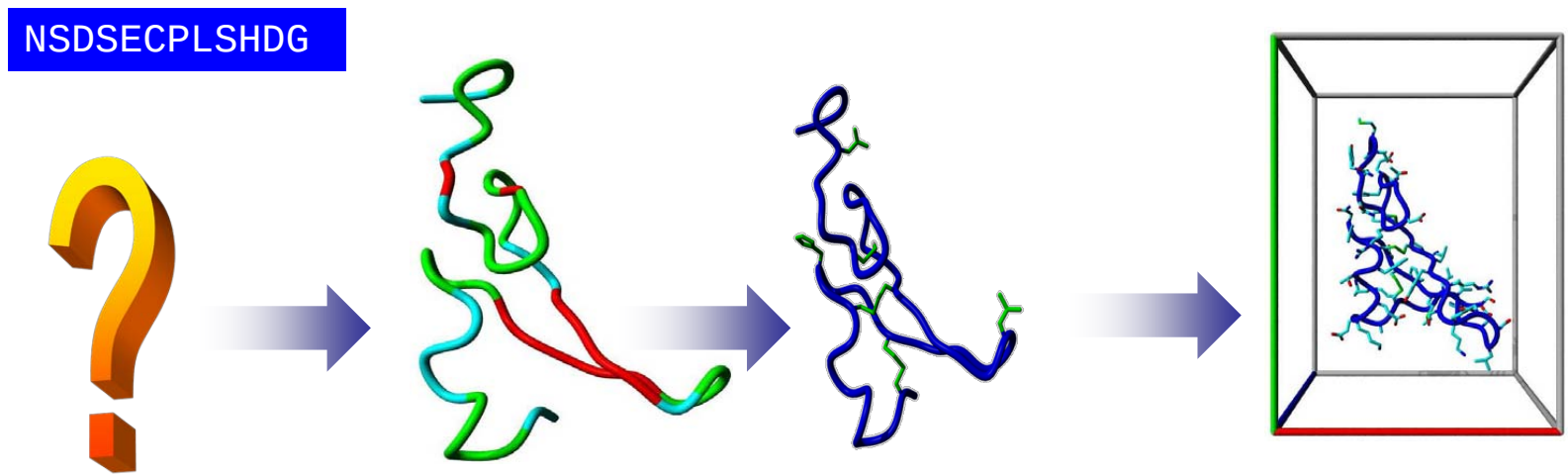
Prediction of structure based upon a highly similar structure



Hanka Venselaar, http://swift.cmbi.ru.nl/teach/B1SEM/B1S_homology_modelling.ppt

Homology modeling in short…

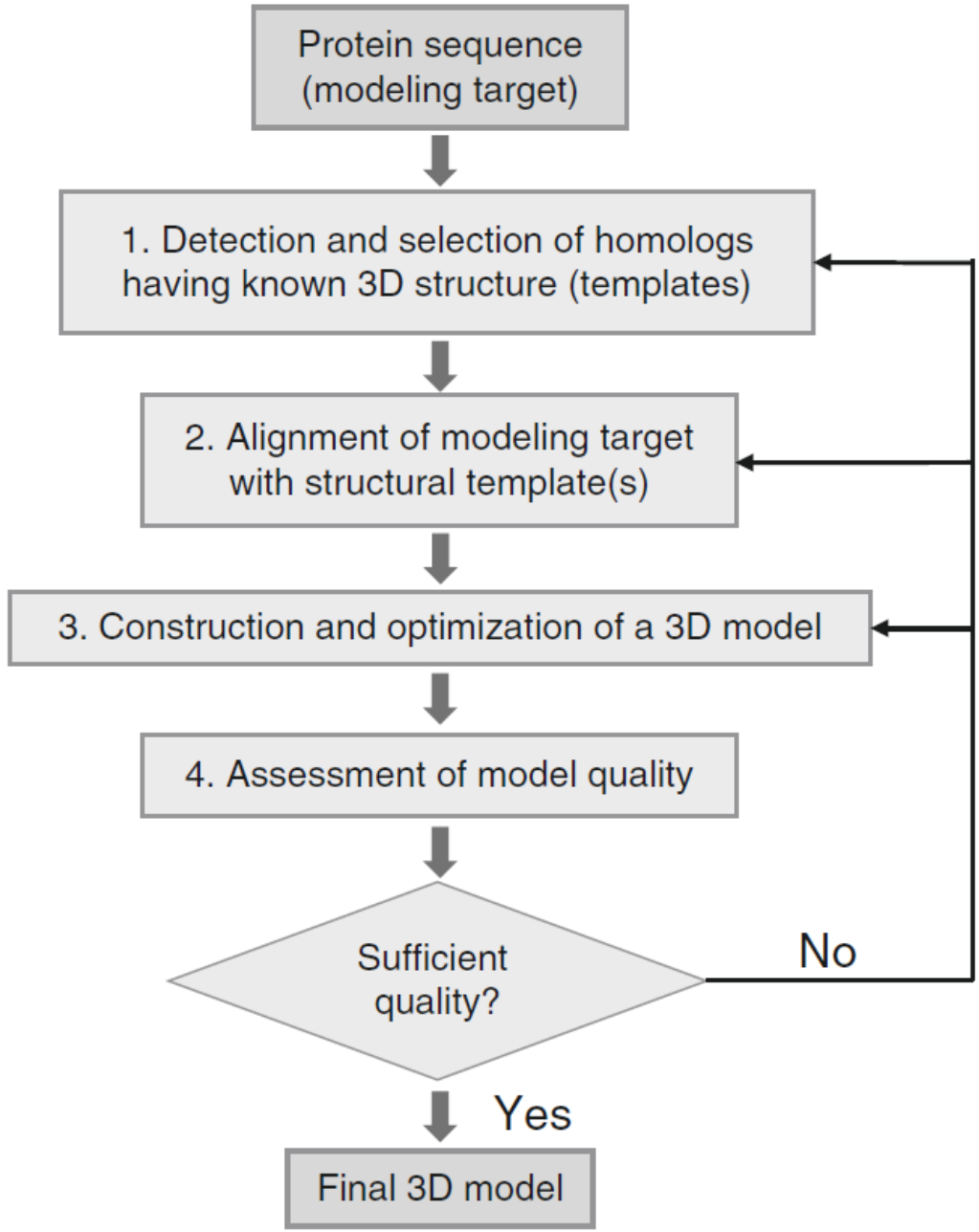
Prediction of structure based upon a highly similar structure



Hanka Venselaar, http://swift.cmbi.ru.nl/teach/B1SEM/B1S_homology_modelling.ppt

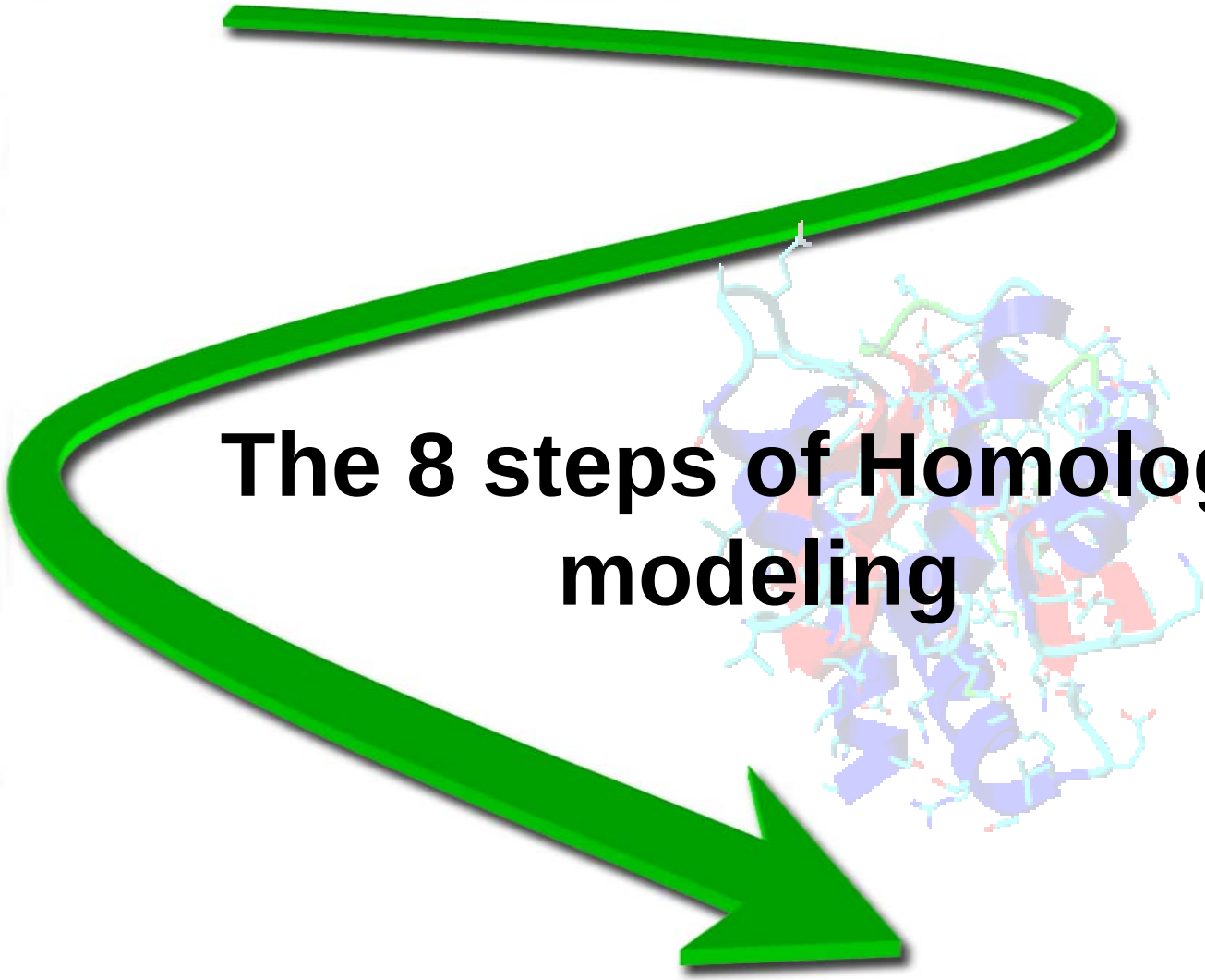
4 grandes étapes

1. Sélection support (template)
2. Alignement cible-template
3. Construction du modèle
4. Évaluation du modèle


Protein sequence
(modeling target)
1. Detection and selection of homologs
having known 3D structure (templates)
2. Alignment of modeling target
with structural template(s)
3. Construction and optimization of a 3D model
4. Assessment of model quality
Sufficient
quality?
No
Yes
Final 3D model

8 petites étapes au final

1. Sélection support (template) et alignement cible-template initial
2. Correction de l'alignement cible-template
3. Construction du modèle: squelette (backbone)
4. Construction du modèle: boucles (loops)
5. Construction du modèle: chaînes latérales (sidechains)
6. Construction du modèle: optimisation du modèle
7. Évaluation du modèle
8. Itération (=reprise) des étapes précédentes



The 8 steps of Homology modeling

Hanka Venselaar, http://swift.cmbi.ru.nl/teach/B1SEM/B1S_homology_modelling.ppt

1: Template recognition and initial alignment

- BLAST your sequence against PDB
- Best hit → normally template
- Initial alignment →

Alignment	DB:ID	Source	Length	Score	Identity%	Positives%	E()
1	PDB:1NQL_A	mol:protein length:624 Epidermal Growth Factor Receptor	624	2728	100	100	0.0
2	PDB:1MOX_B	mol:protein length:501 Epidermal Growth Factor Receptor	501	2728	100	100	0.0
3	PDB:1MOX_A	mol:protein length:501 Epidermal Growth Factor Receptor	501	2728	100	100	0.0
4	PDB:1IVO_B	mol:protein length:622 Epidermal Growth Factor Receptor	622	2728	100	100	0.0


```
NSDSECPLSHDGYCLHDGVC
||   || | |||||  |||
NSYPGCPSSYDGYCLNGGVC
```

2: Alignment correction

- Functional residues → conserved
- Use multiple sequence alignments
- Deletions → shift gaps

```
CPISRTAAS-FRCW
CPISRTG-SMFRCW
CPISRTA--TFRCW
CPISRTAASHFRCW
CPISRTGASIFRCW
CPISRTA---FRCW
```

Multipe sequence alignment

```
CPISRTGASIFRCW    CPISRTGASIFRCW
CPISRTA---FRCW    CPISRT---AFRCW
```

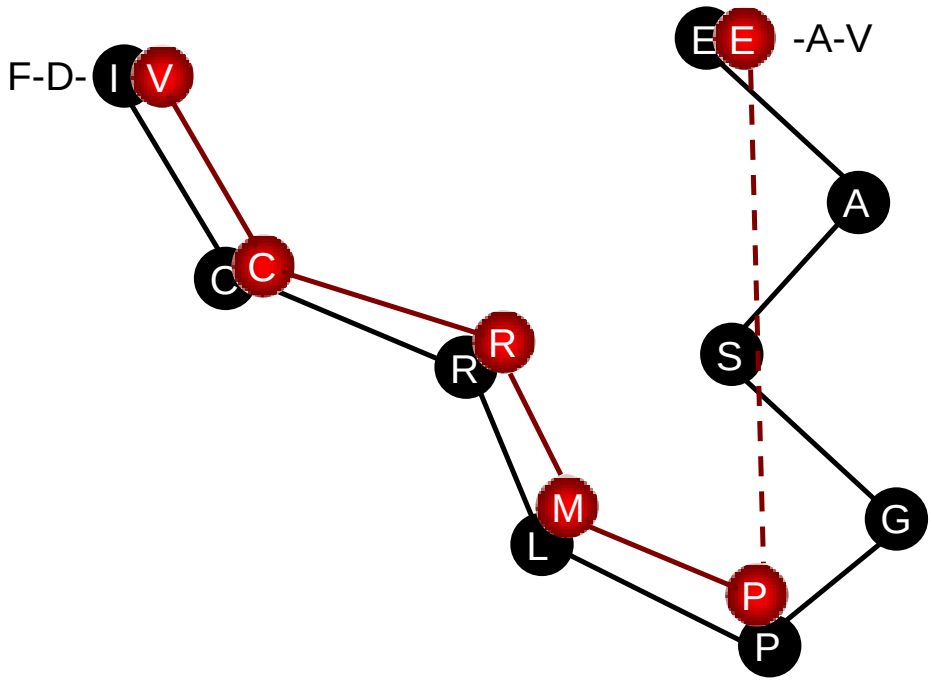
← Sequence with known structure

←Your sequence

Correct alignment

Hanka Venselaar, `http://swift.cmbi.ru.nl/teach/B1SEM/B1S_homology_modelling.ppt`

2: Alignment correction

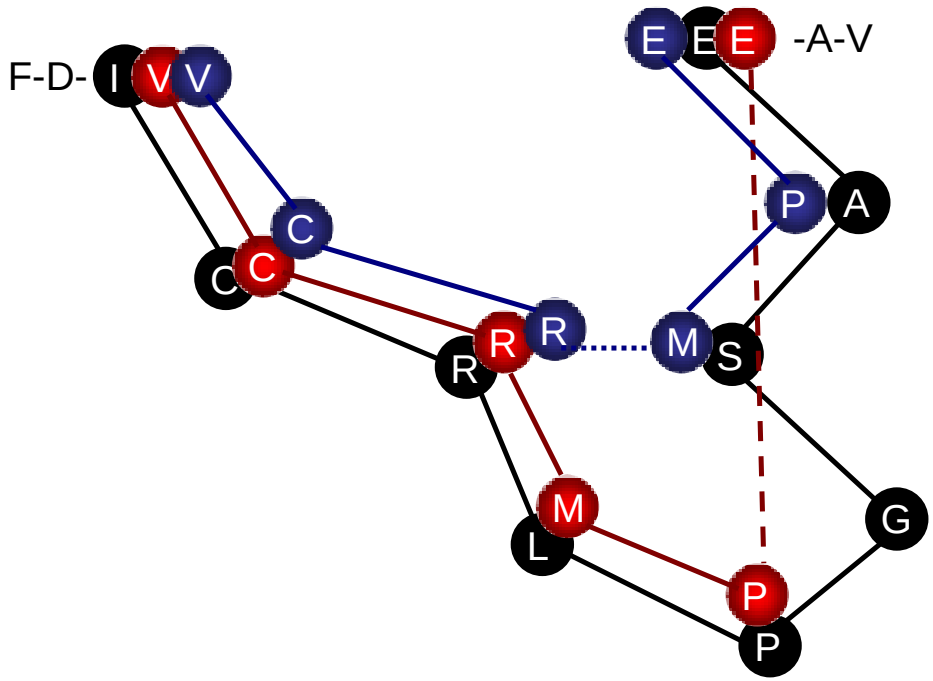


Known structure FDICRLPGSAEAV

Model FNVCRMP---EAI

Hanka Venselaar, http://swift.cmbi.ru.nl/teach/B1SEM/B1S_homology_modelling.ppt

2: Alignment correction



Known structure FDICRLPGSAEAV

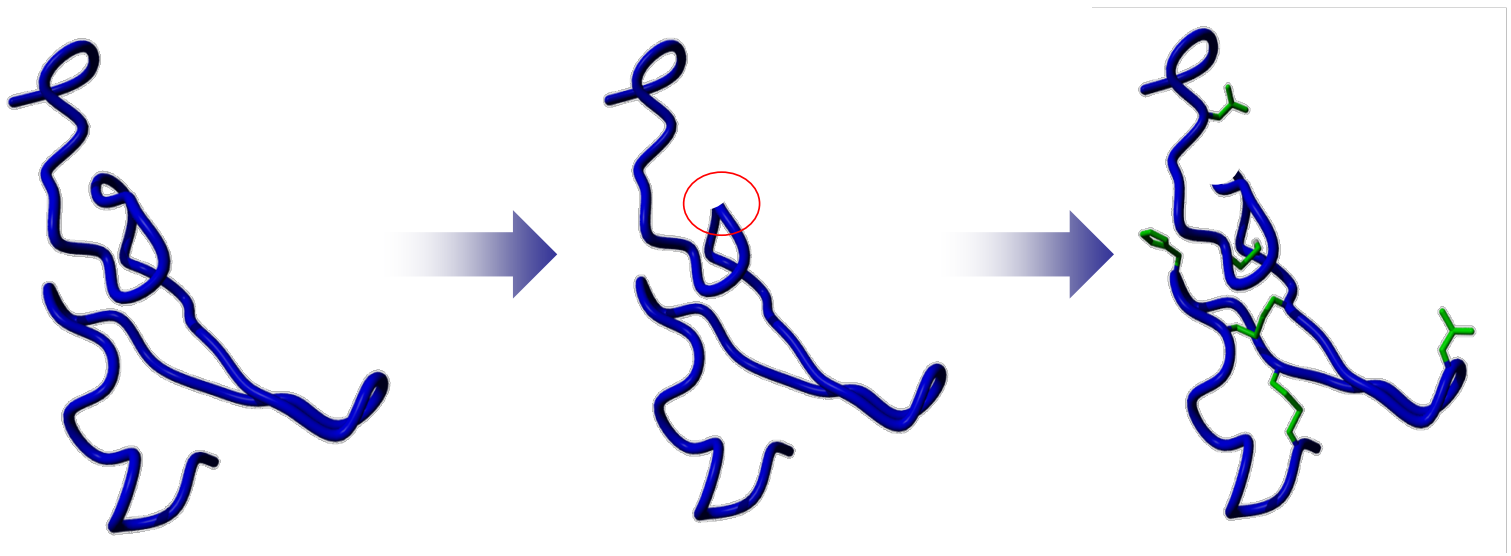
Model FNVCRMP---EAI

Model FNVCR---M**P**EAI ← Correct alignment

Hanka Venselaar, http://swift.cmbi.ru.nl/teach/B1SEM/B1S_homology_modelling.ppt

3: Backbone generation

- Making the model….
- Copy backbone of template to model
- Make deletions as discussed
- (Keep conserved residues)

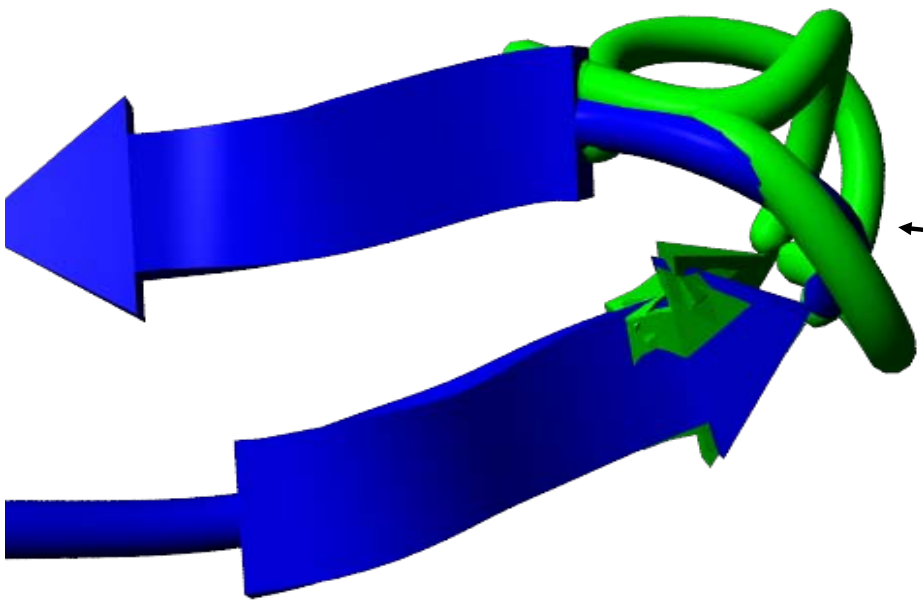


Hanka Venselaar, `http://swift.cmbi.ru.nl/teach/B1SEM/B1S_homology_modelling.ppt`

4: Loop modeling

Known structure **GVCMYIEA---LDKYACNC**

Your sequence **GECFMVKDLSNPSRYLCKC**



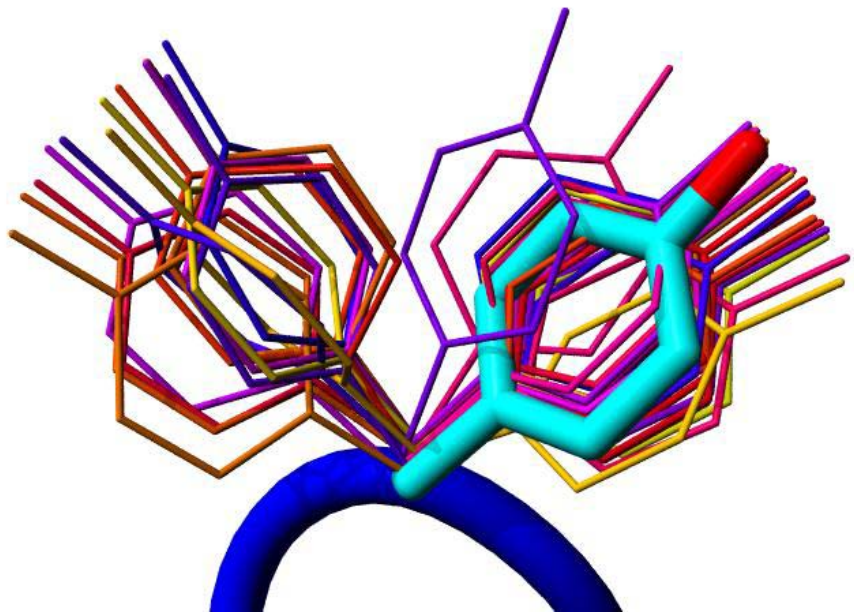
Loop library,

try different options

Hanka Venselaar, http://swift.cmbi.ru.nl/teach/B1SEM/B1S_homology_modelling.ppt

5: Side-chain modeling

- Several options
- Libraries of preferred rotamers based upon backbone conformation

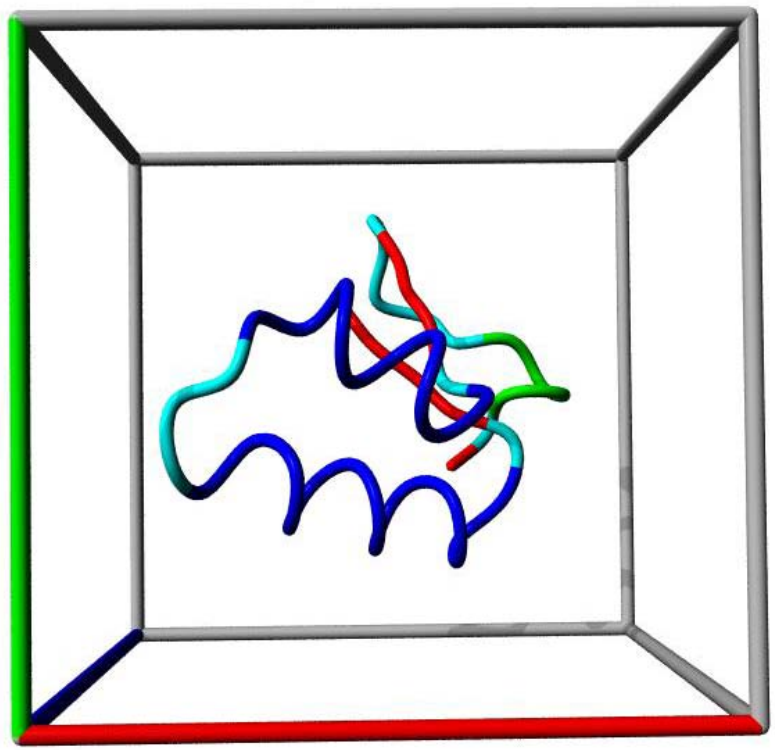


Hanka Venselaar, http://swift.cmbi.ru.nl/teach/B1SEM/B1S_homology_modelling.ppt

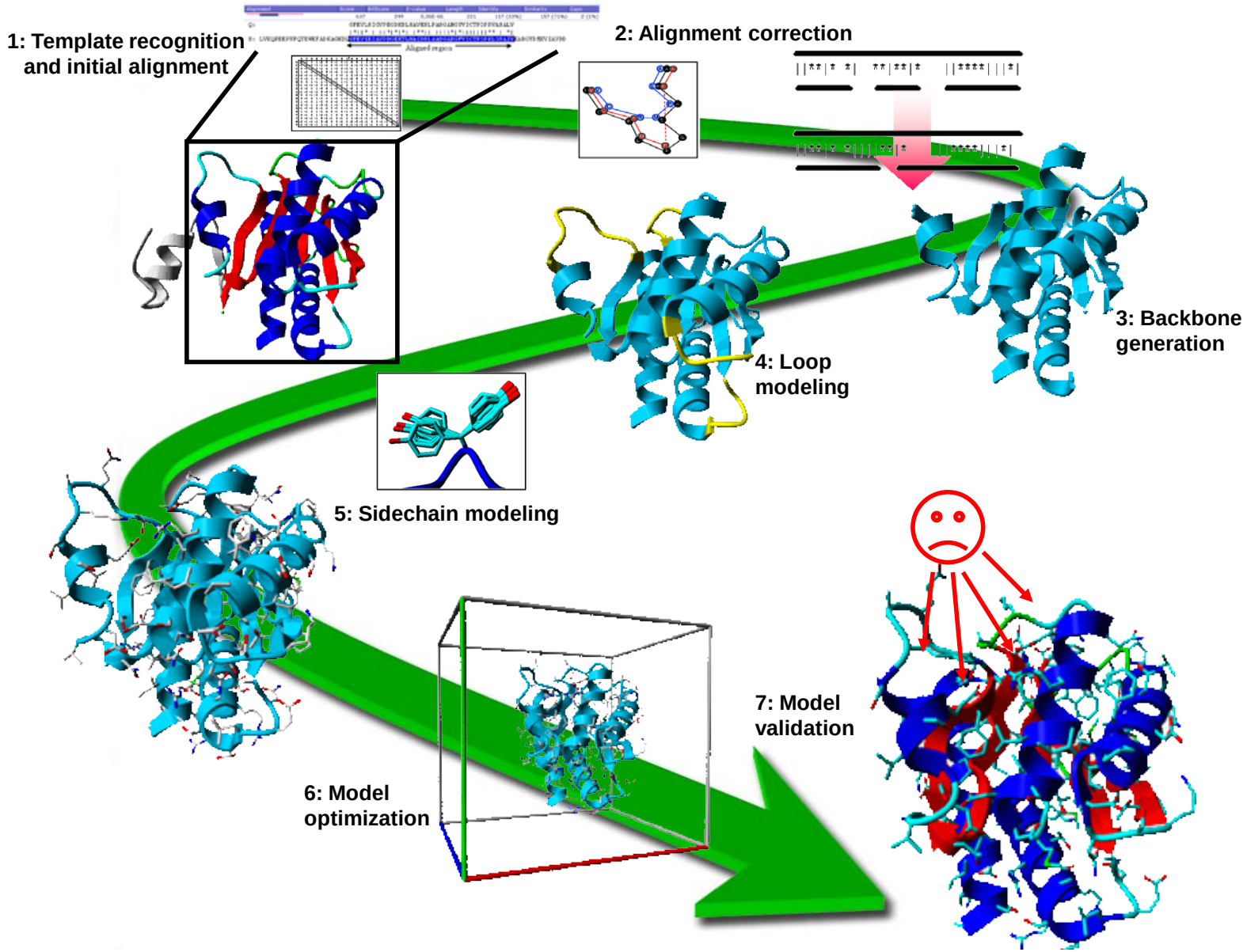
6: Model optimization

- Molecular dynamics simulation
- Remove big errors

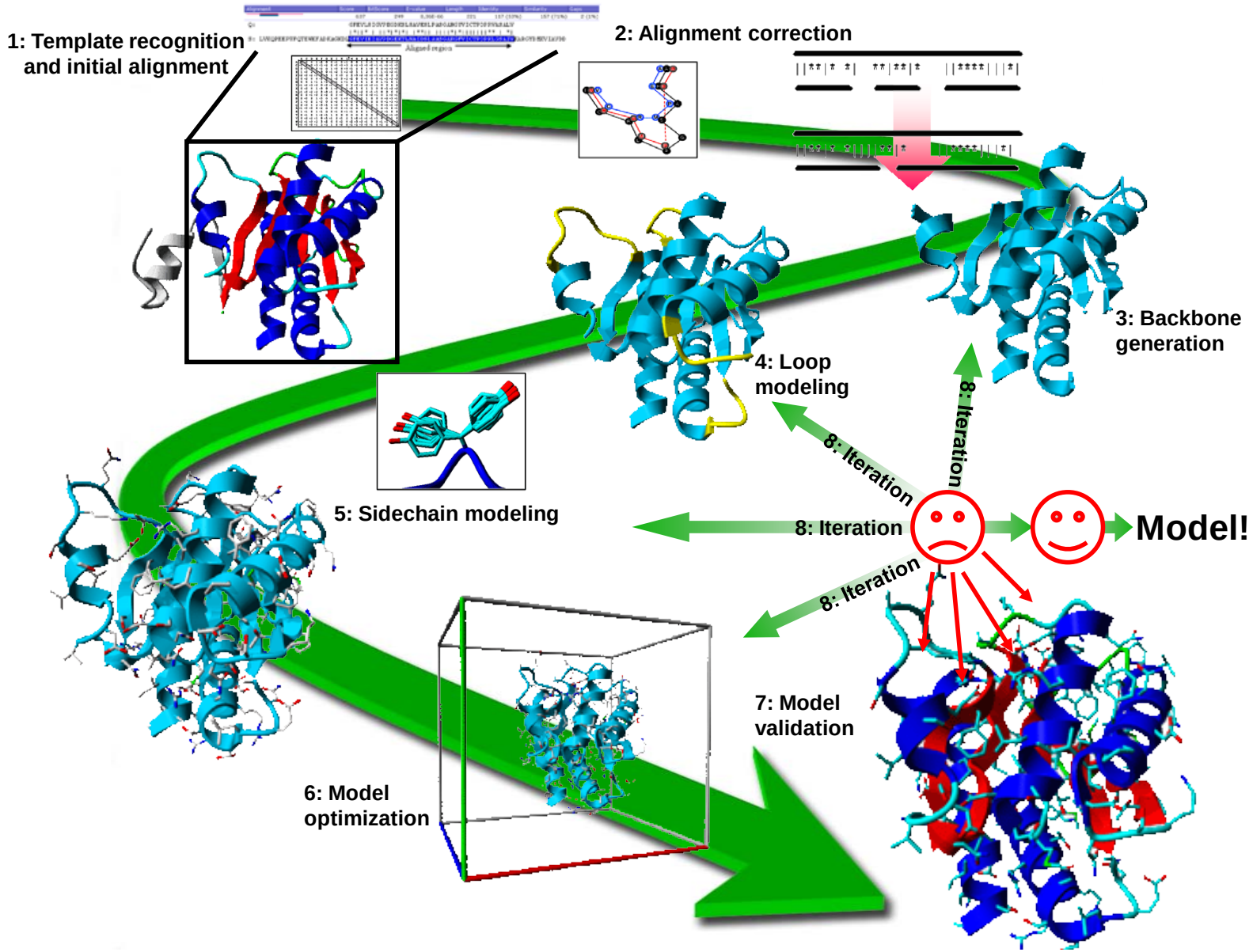
- Structure moves
  to lowest
  energy conformation



Hanka Venselaar, http://swift.cmbi.ru.nl/teach/B1SEM/B1S_homology_modelling.ppt



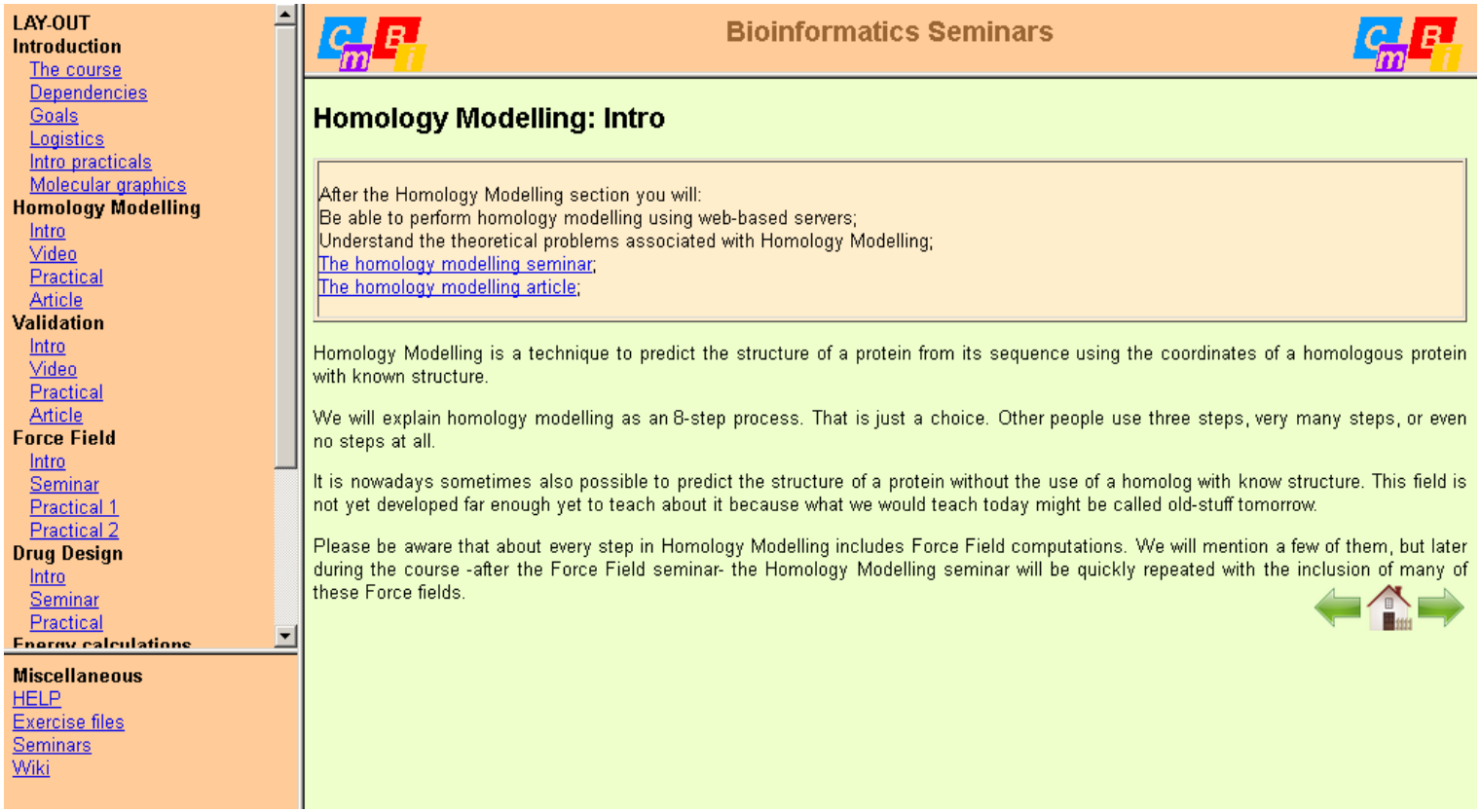
Hanka Venselaar, `http://swift.cmbi.ru.nl/teach/B1SEM/B1S_homology_modelling.ppt`



Hanka Venselaar, `http://swift.cmbi.ru.nl/teach/B1SEM/B1S_homology_modelling.ppt`

CMBI courses

http://swift.cmbi.ru.nl/teach/B1SEM/



Article:

http://swift.cmbi.ru.nl/teach/B1SEM/HTML/hanka_modelling.pdf

Conditions

1. Existence d'un template
2. Identité de séquence cible <-> template > XX %
3. Alignement correct entre séquence cible et structure 3D du template

Sélection du template

Introduction

- La sélection du template et l'alignement cible - template sont les deux étapes clés de la modélisation par homologie, car un mauvais choix du template ou un mauvais alignement ne peut pas plus être corrigé par la suite.
- Les deux étapes vont de pair
- Des méthodes, comme HHpred/HHsearch, qui sont spécialisée dans ces deux étapes n'ont pas besoin d'une construction de modèle sophistiquée pour être parmi les premiers dans CASP

Outils d'annotation

template = support 3D ou patron
Annotation (séquence, structure, fonction):

- InterProScan: Identification domaines, motifs, familles
- PsiPred: Prédiction structure secondaire
- DisoPred: Prédiction désordre (=> IUP)
- MEMSAT: Prédiction ségments transmembranaires
- Choix du template doit être fait manuellement pour choisir les structures qui correspondent au même état fonctionel que la cible.
- Possible de choisir plusieurs templates en même temps pour mieux couvrir la séquence de la cible

Alignement cible - template

Étapes

template = support 3D ou patron

1. alignement initial séquence-structure <=> choix des templates
2. trouver les zones d'alignement qui demandent un ajustement
3. amélioration de l'alignement

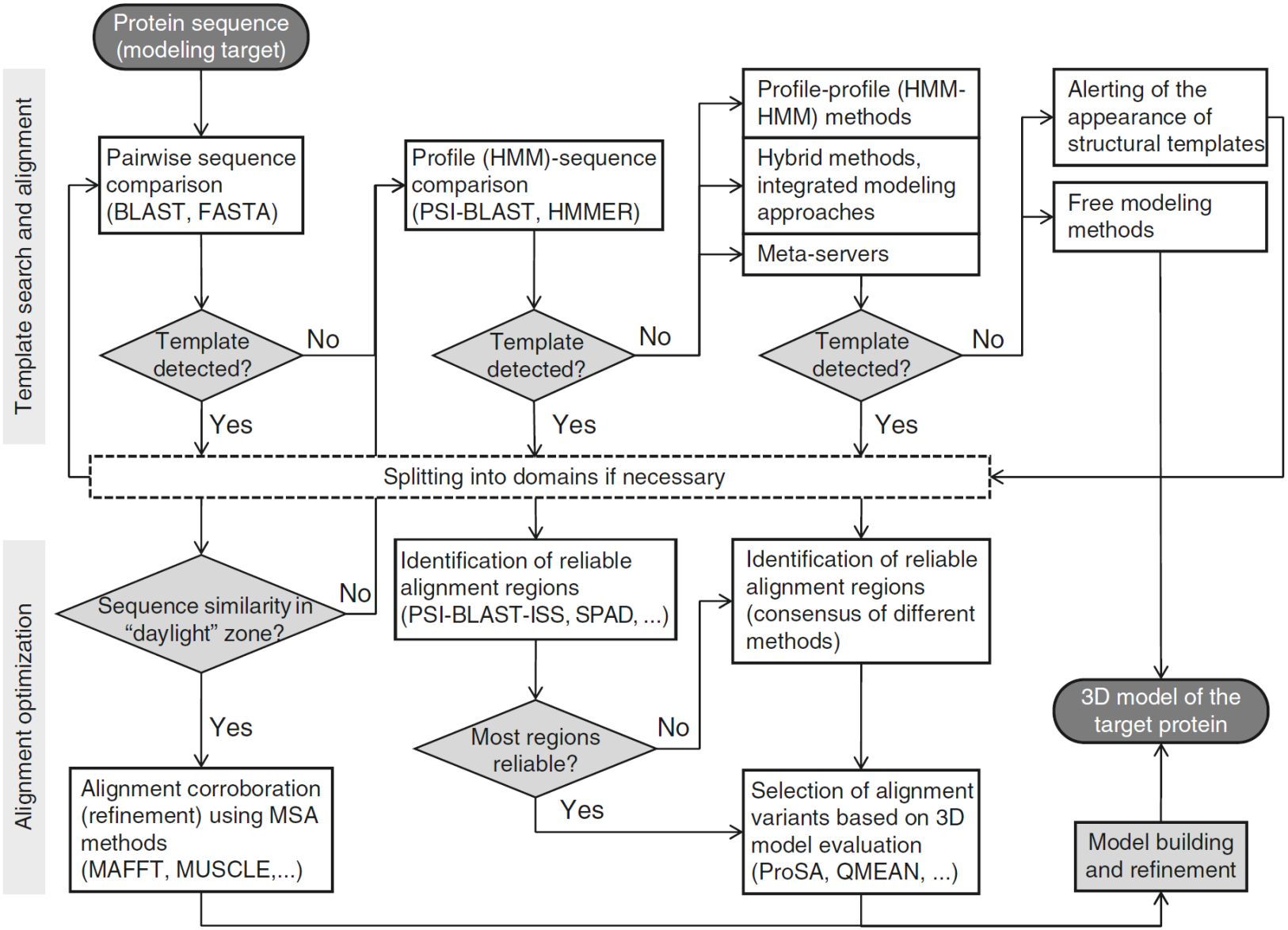


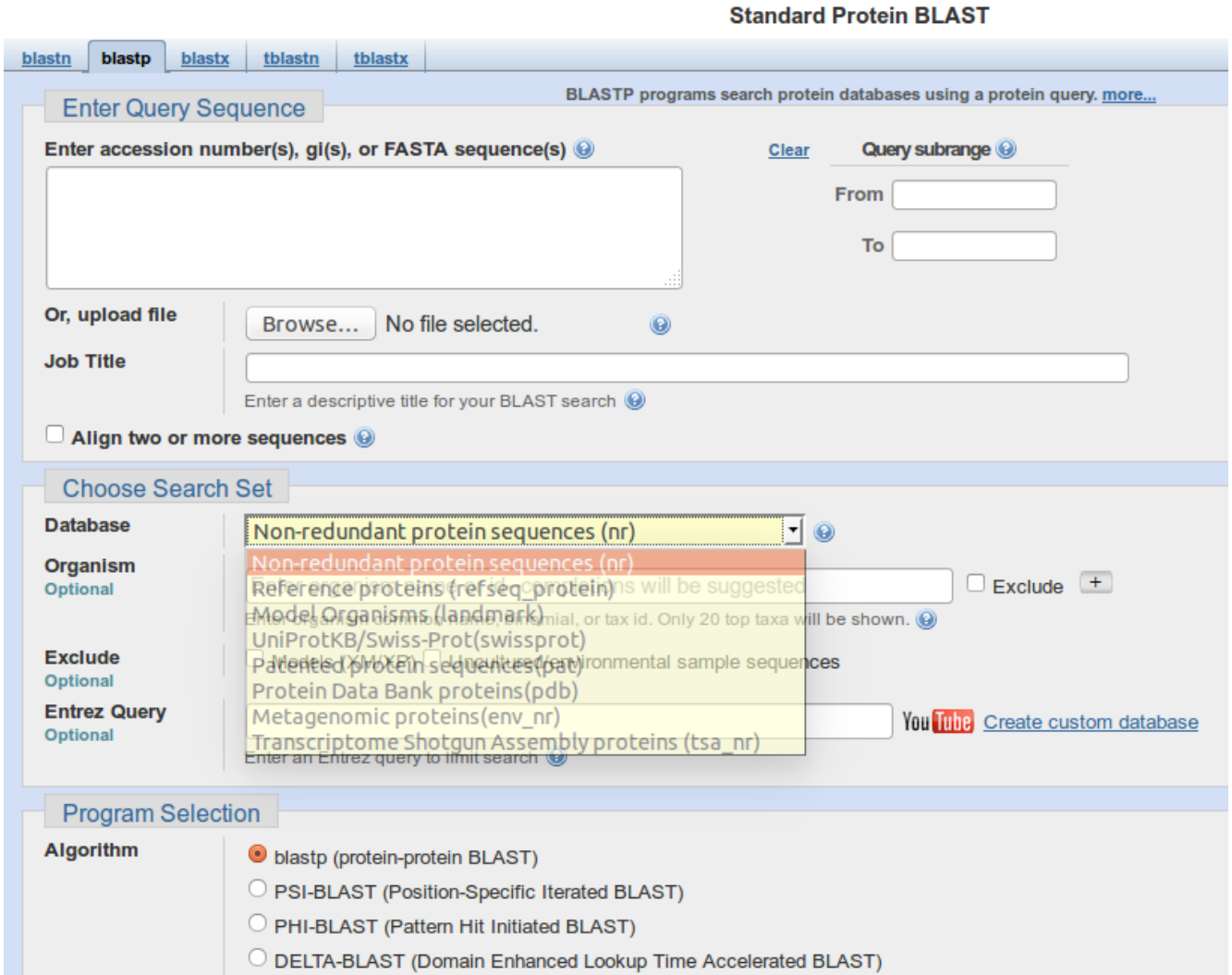
Fig. 6. Flowchart of major steps in sequence to structure alignment.

Andreeva, "Homology Modeling", ch. 3, Methods in Mol. Biol.(2012)

BLAST

- suffit si identité de séquence > 40% et statistiquement significatif (E value < 0.001 (expectation value))
- mais il est quand même recommandé d'assembler un jeu de séquences homologues avec BLAST puis de les aligner avec une méthode MSA

Attention au choix de la database pour BLAST



Databases pour BLAST

- BLAST database: "Non-redundant protein sequences" ("nr"):
  20 millions de séquences le 29/09/2012
  100 millions de séquences le 27/09/2016
  238 millions de séquences le 22/03/2019
  509 millions de séquences le 06/10/2022
  480k séquences SwissProt 142k séquences PDB
  `http://blast.ncbi.nlm.nih.gov`
- PDB database:
  60091 structures 3D le 29/09/2012
  86698 structures 3D le 27/09/2016
  109709 structures 3D le 07/04/2019
  196108 structures 3D le 06/10/2022
  et 1 millions "computed structure models (CSM)"

PSI-BLAST

- Le nombre de séquences dans la PDB est trop petit pour appliquer PSI-BLAST directement
- Procédure "PDB-BLAST":
  1. plusieurs itérations avec PSI-BLAST avec la base "nr"
  2. utiliser le profil construit pour faire une dernière itération avec les séquences de la PDB
  3. Un serveur: `http://protein.bio.unipd.it/pdbblast/`

Séparation en domaines structurales

- Si la cible est une protéine multi-domaine, il est nécessaire de la séparer en ses domaines structurales
- Les domaines seuls peuvent être plus proche par rapport aux templates

Pas de template :-(

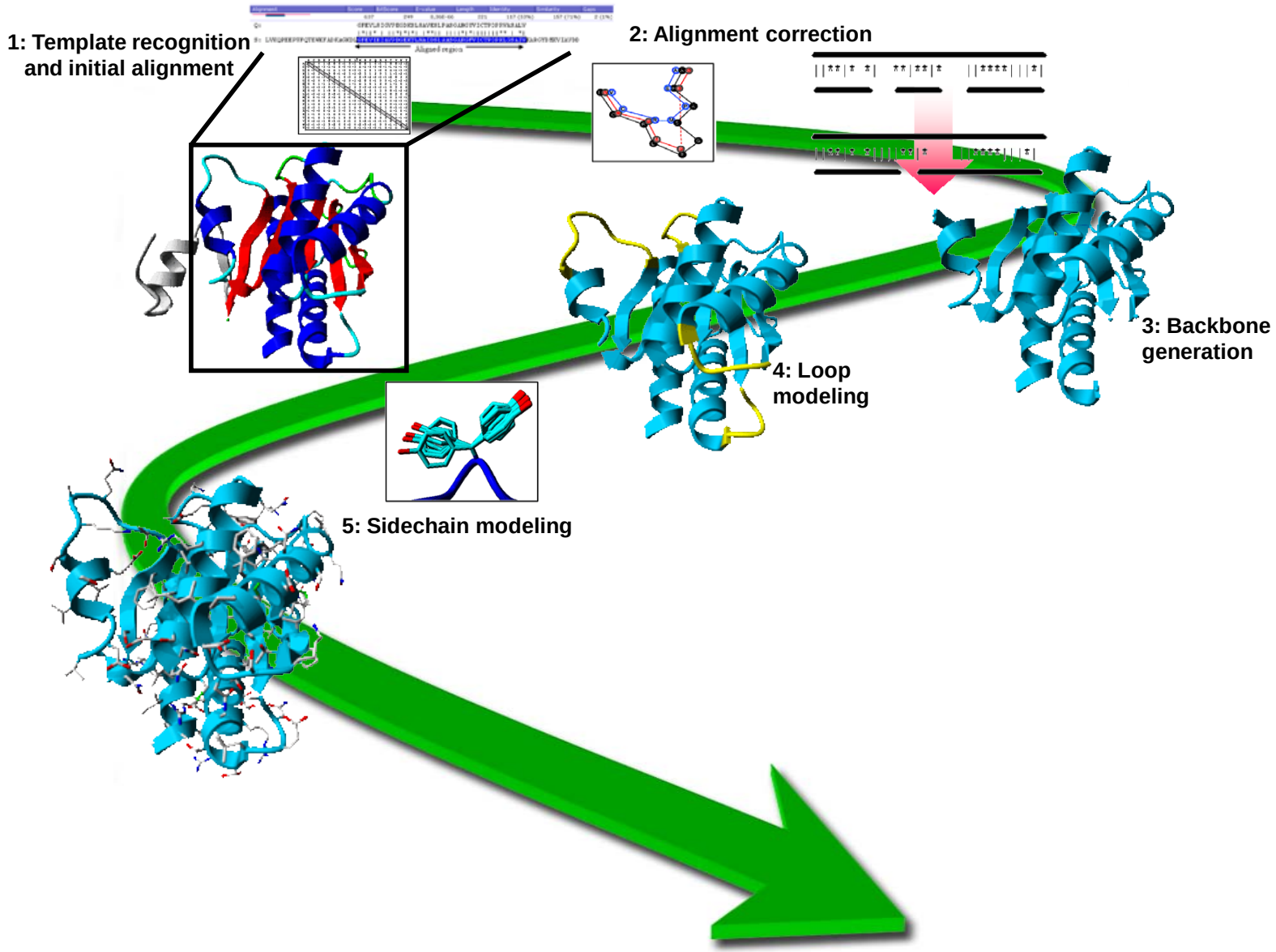
Si il n'y a vraiment pas de template dans la PDB, alors il reste comme options:

1. Si ce n'est pas urgent, attendre qu'un template est ajouté à la PDB.
   Système automatique d'alerte: PDBalert (utilise HHsearch, très sensible)
2. Free Modeling (FM):
   - Robetta (serveur basé sur Rosetta)
   - I-TASSER
   - SAM-T08
   - MULTICOM
3. FM marche raisonnablement pour des protéines de petite taille (< 100 résidus) et de topologie simple

Construction du modèle

Assemblage de corps rigides

- Assemblage des templates en tant que corps rigide
- Basé sur la décomposition naturel de la protéine en:
  1. Coeur/Noyau conservé
  2. Boucles variables qui lient les régions conservé
  3. Chaînes latérales qui décorent le squelette



Hanka Venselaar, `http://swift.cmbi.ru.nl/teach/B1SEM/B1S_homology_modelling.ppt`

Assemblage de corps rigides

Exemple avec COMPOSER:

1. Templates sont superposées
2. *Framework* = Moyenne des coordonnées Cα des régions structurellement conservées
3. Superposer sur le framework le coeur du meilleur template comme début pour le modèle
4. Générations des boucles par une recherche dans une banque de structure 3D de boucles
5. Chaînes latérales sont modélisées avec leur préférences conformationnelles intrinsèque
6. Affinement par minimisation d'énergie ou dynamique moléculaire

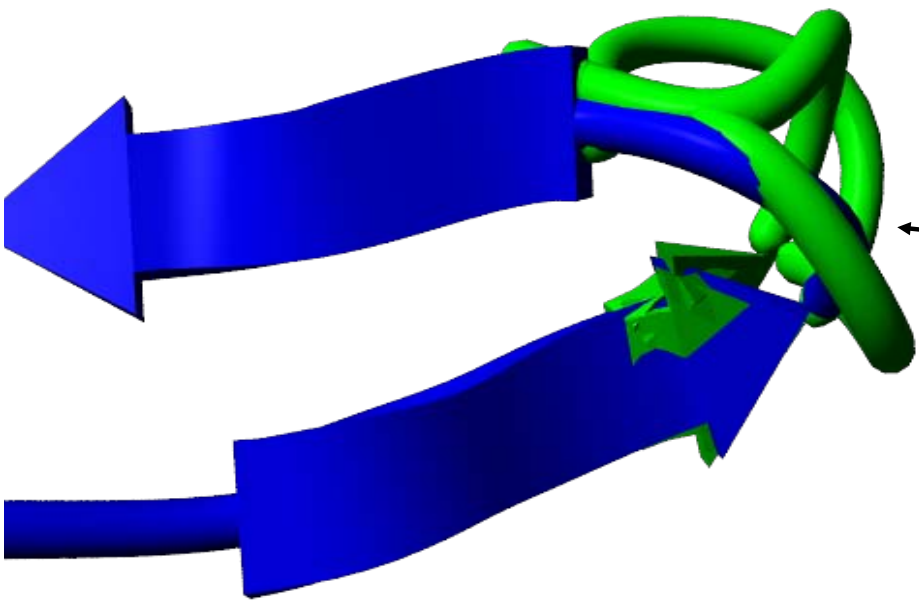
Modélisation par satisfaction de contraintes spatiales

- Concept similaire que pour la détermination d'une structure par RMN
- Contraintes de distances obtenues depuis le template
- Application d'un champ de force de mécanique moléculaire
- Modèle doit satisfaire le plus de contraintes que possible
- Exemple: MODELLER

4: Loop modeling

Known structure **GVCMYIEA---LDKYACNC**

Your sequence **GECFMVKDLSNPSRYLCKC**



Loop library,

try different options

Hanka Venselaar, http://swift.cmbi.ru.nl/teach/B1SEM/B1S_homology_modelling.ppt

Modélisation des boucles

- Souvent: l'insertion de boucles à la surface
- Ces boucles sont importants pour la liaison de ligands
- Une modélisation correcte de ses boucles est important pour pouvoir utiliser le modèle par la suite dans des études de docking de ligand.
- Mini protein folding problem:
  On ne connaît que la séquence.
  Par contre les boucles sont trop courts pour avoir assez d'information sur leur conformation.
- Points d'ancrage sur la structure
- Difficile de modéliser des boucles au delà de 8 résidus
- La plupart des insertions sont plus courtes que 10-12 résidus

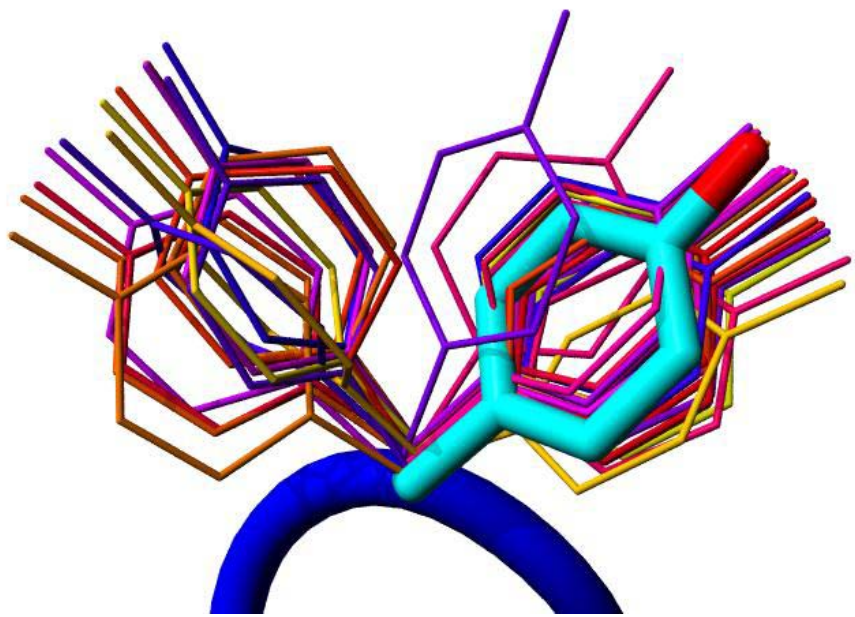
Modélisation des boucles

Deux approches:

1. Recherche dans la PDB pour trouver des segments qui sont compatibles avec les points d'ancrage.
   Limites:
   - Le nombre de conformations possibles croit exponentiellement avec la longueur de la boucle.
   - Limitation en taille de boucle < 9 résidus
2. Recherche conformationnelle: optimisation avec score
   Exemple: MODELLER: gradient conjugé et dynamique moléculaire avec recuit simulé

5: Side-chain modeling

- Several options
- Libraries of preferred rotamers based upon backbone conformation



Hanka Venselaar, http://swift.cmbi.ru.nl/teach/B1SEM/B1S_homology_modelling.ppt

Modélisation des chaînes latérales

Observations:

1. Un échange d'acides aminés laisse le squelette souvent inchangé
   => Squelette rigide lors de la recherche des conformations des chaînes latérales
2. Nombre de conformations est limité (contraintes stéréochimiques et énergétiques)
   => Banque de rotamères des chaînes latérales
3. Limitation: le score et non l'échantillonnage

Conformations chaînes latérales

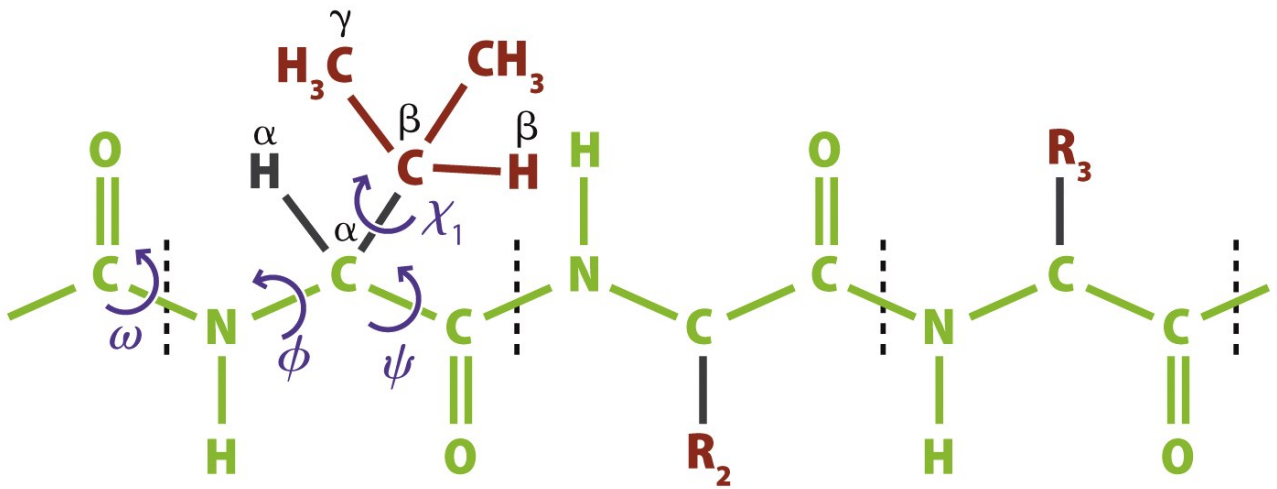


Figure 1.4 How Proteins Work (©2012 Garland Science)

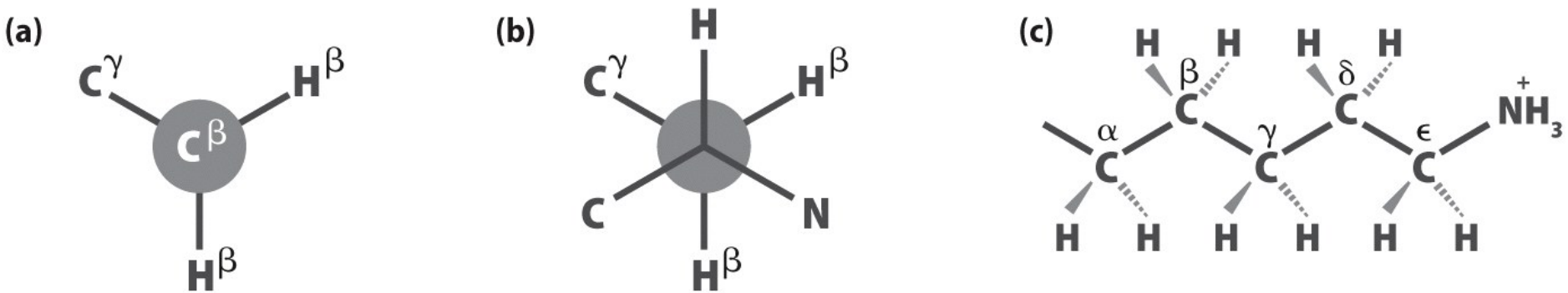
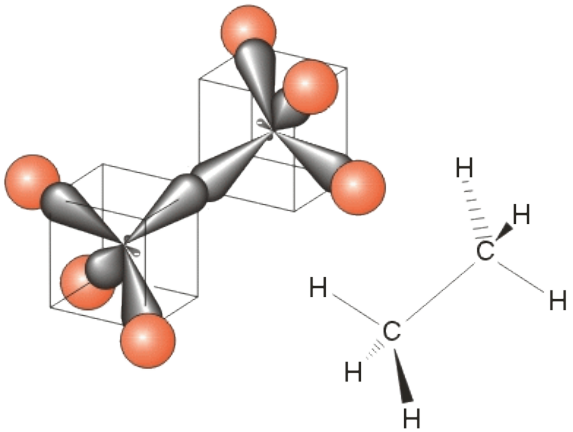
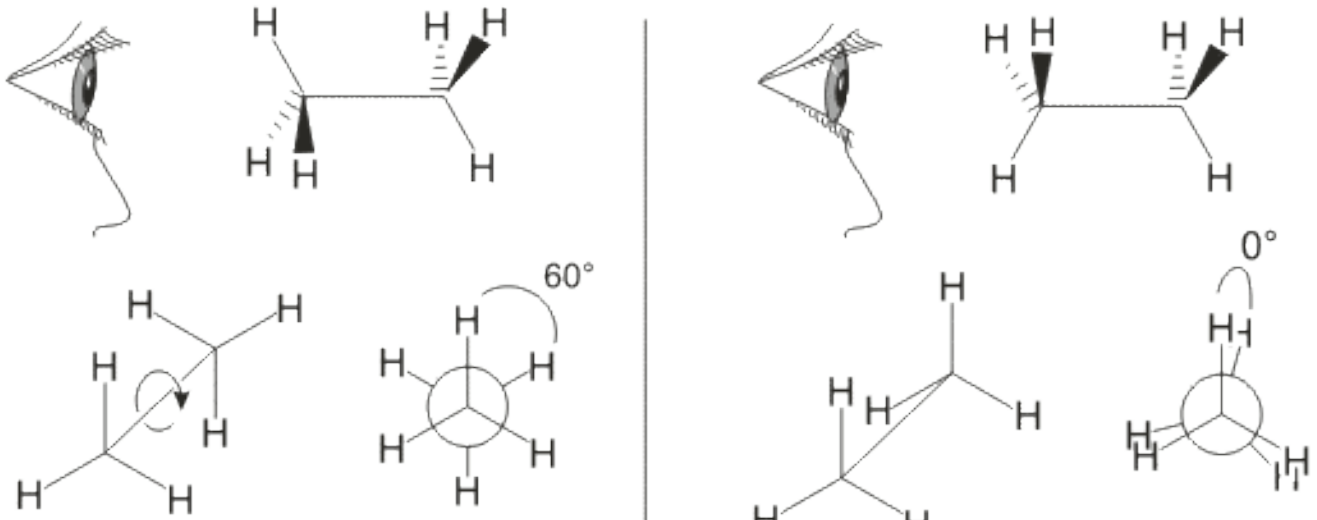


Figure 1.9 How Proteins Work (©2012 Garland Science)

Rotamères de l'éthane

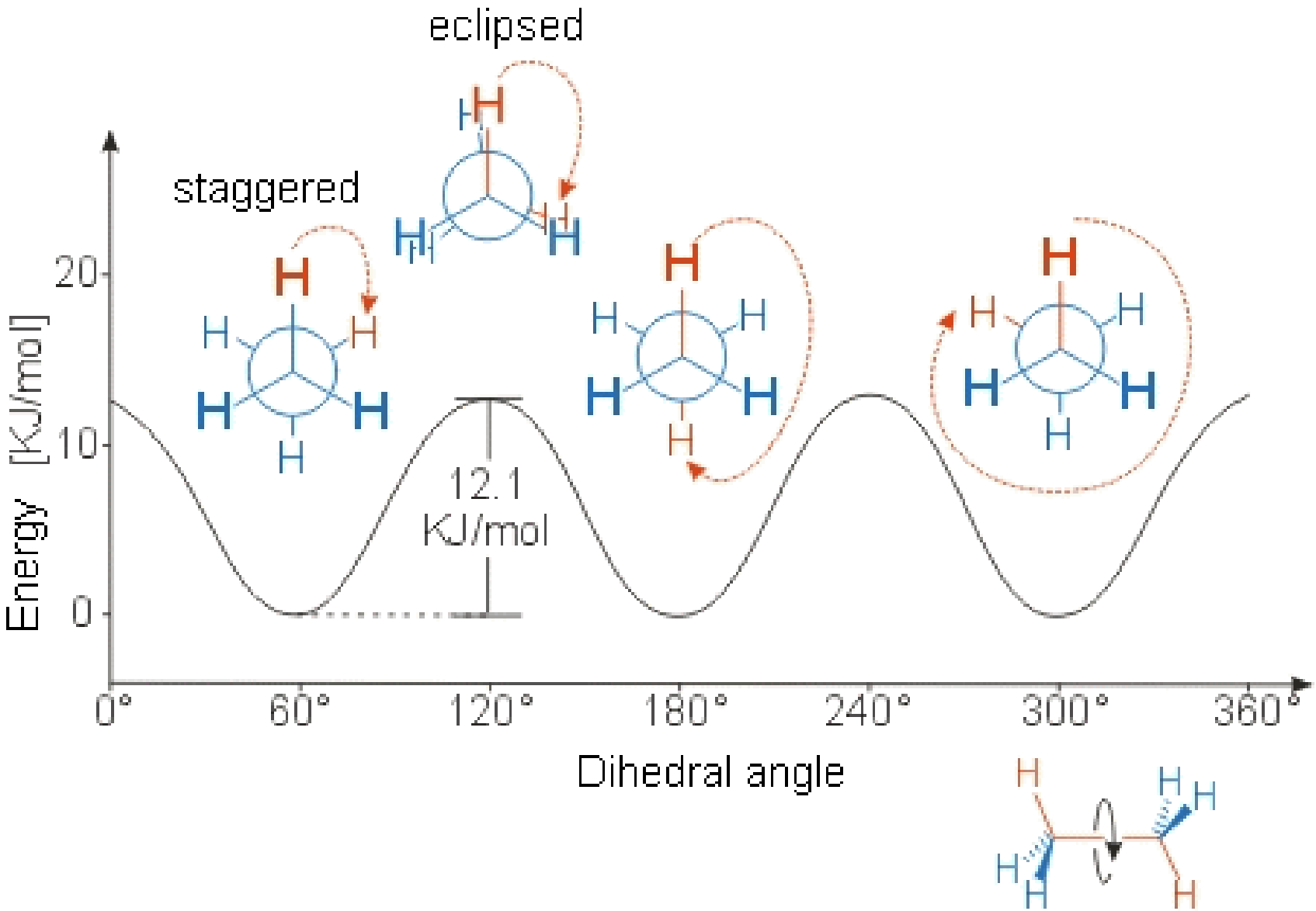




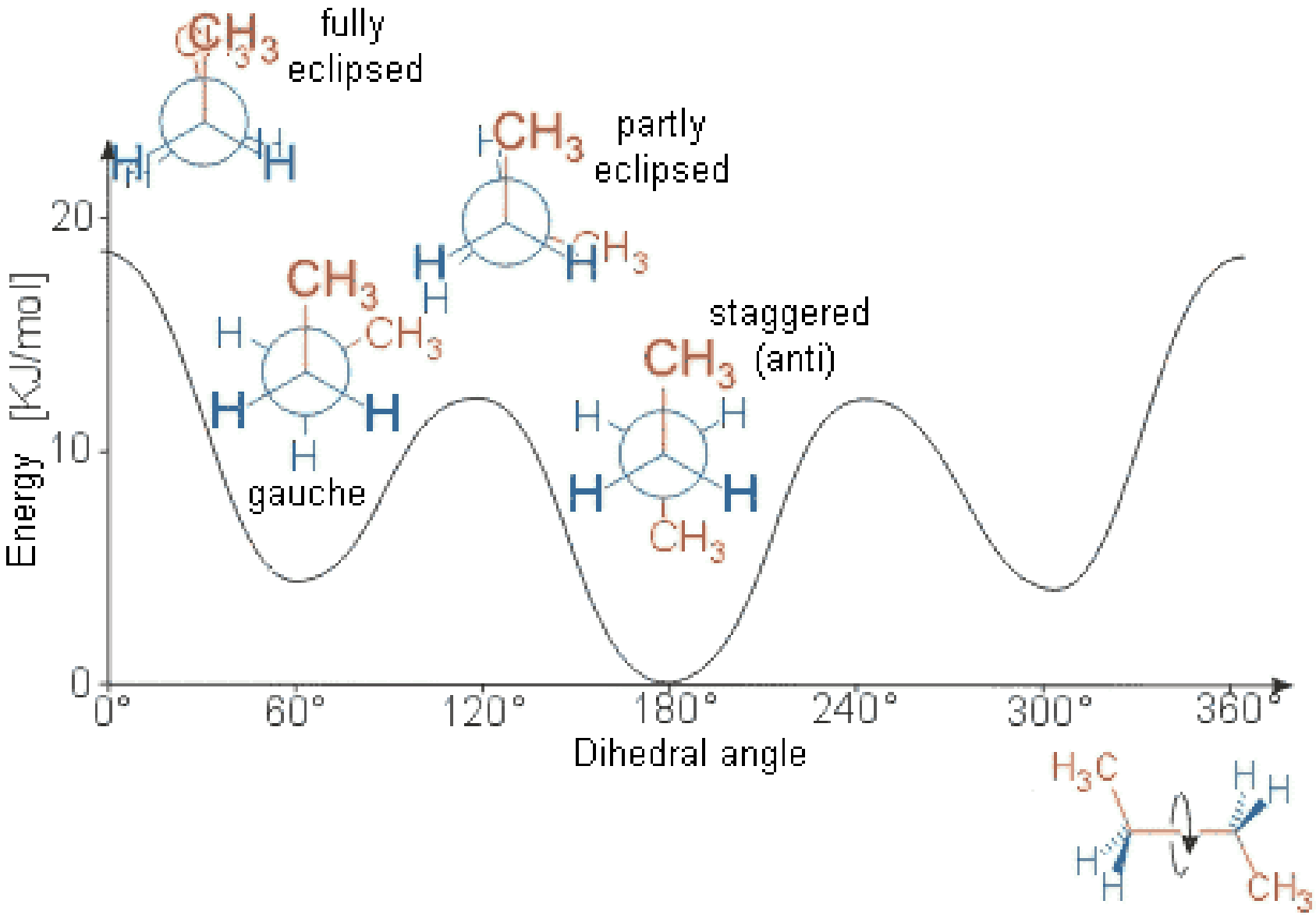
Sawhorse Newman Sawhorse Newman

Staggered Eclipsed

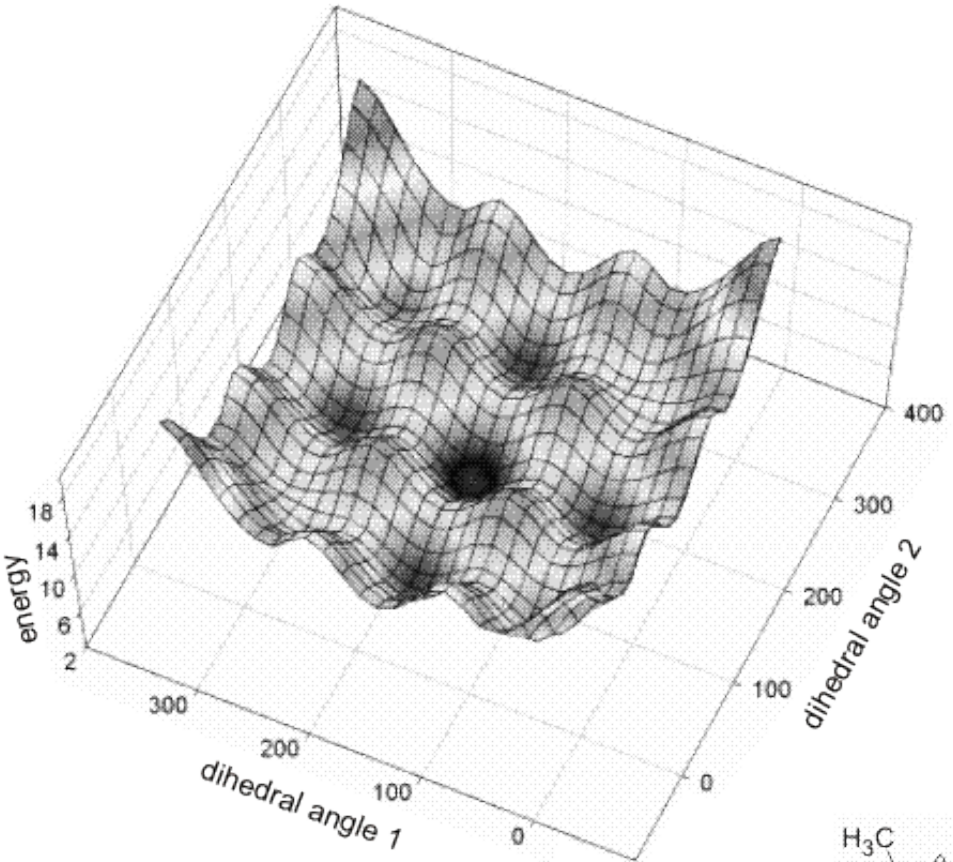
Rotamères de l'éthane

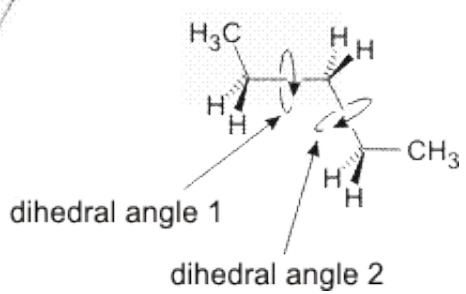


Rotamères du butane



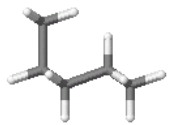
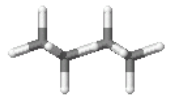
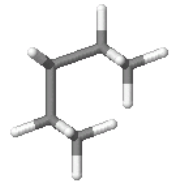
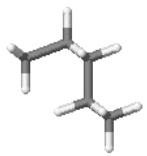
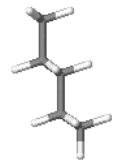
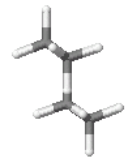
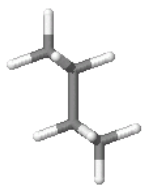
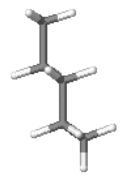
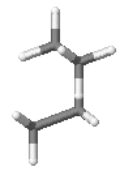
Rotamères de pentane





3D Models and properties of the energy minima of pentane

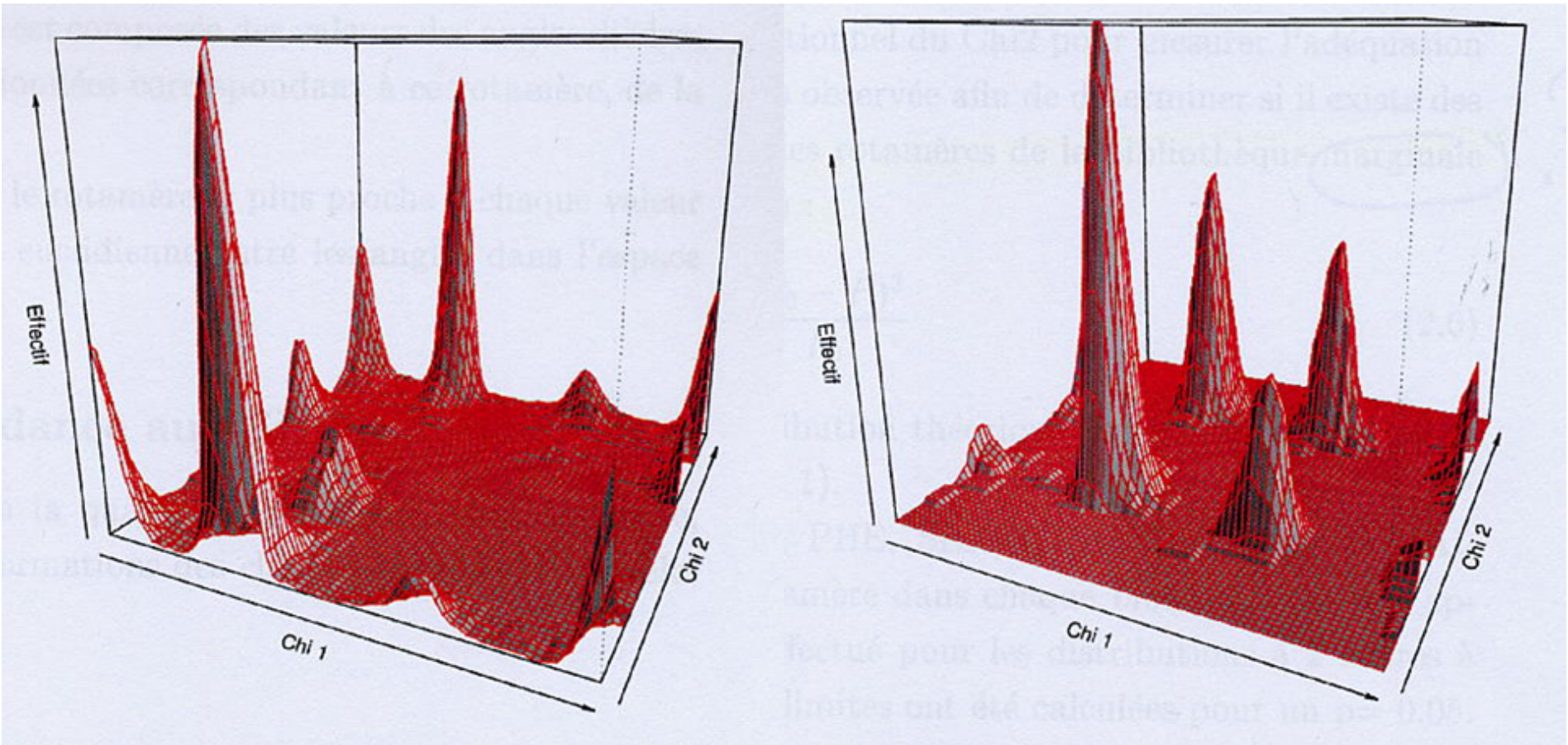
Tab.1 | **Conformations of energy minima**

	 Jmol	 Jmol	 Jmol
	g^-g^+ $\theta_1 = 300°$ $\theta_2 = 300°$ C_2 symmetry	ag^+ $\theta_1 = 180°$ $\theta_2 = 300°$	g^+g^+ $\theta_1 = 60°$ $\theta_2 = 300°$ C_S symmetry
	 Jmol	 Jmol	 Jmol
	g^-a $\theta_1 = 300°$ $\theta_2 = 180°$	aa $\theta_1 = 180°$ $\theta_2 = 180°$ C_{2v} symmetry	g^+a $\theta_1 = 60°$ $\theta_2 = 180°$
	 Jmol	 Jmol	 Jmol
	g^-g^- $\theta_1 = 300°$ $\theta_2 = 60°$ C_S symmetry	ag^- $\theta_1 = 180°$ $\theta_2 = 60°$	g^+g^- $\theta_1 = 60°$ $\theta_2 = 60°$ C_2 symmetry

Rotamères

ARG

PHE



Rotamères des acides aminés

Table I
Side-Chain Dihedral Angles Used as Starting Points for Energy Minimization[a]

Residue	Side-chain dihedral angles, deg			
	χ^1	$\chi^{2\,b}$	$\chi^{3\,c}$	$\chi^{4\,d}$
Ala	60			
Arg	±60, 180	±60, 180	±60, 180	±60, 180[e]
Asn	±60, 180	±30, ±90, ±150	180	
Asp	±60, 180	±30, ±90, ±150	0, 180	
Cys	±60, 180	±60, 180		
Gln	±60, 180	±60, 180	±30, ±90, ±150	180
Glu	±60, 180	±60, 180	±30, ±90, ±150	0, 180
Gly	*f*			
His[g]	±60, 180	±30, ±90, ±150		
Ile	±60, 180	±60, 180	60	60
Leu	±60, 180	±60, 180	60	60
Lys	±60, 180	±60, 180	±60, 180	±60, 180[h]
Met	±60, 180	±60, 180	±60, 180	60
Phe	±60, 180	90, ±30		
Pro	*f*			
Ser	±60, 180	±60, 180		
Thr	±60, 180	±60, 180	60	
Trp	±60, 180	±30, ±90, ±150		
Tyr	±60, 180	±30, ±90, ±150	0	
Val	±60, 180	60	60	

3. Modélisation similitude: minim énergie

- Dernière étape de la modélisation moléculaire
- Suppression de l 'encombrement stérique des chaînes latérales
- Mise en conformité avec les valeurs canoniques des angles de valence, des dièdres....

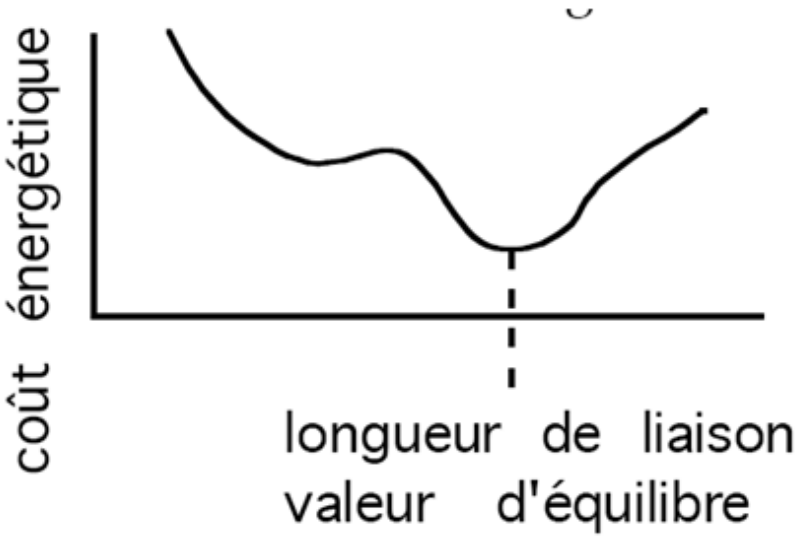
Minimisation

L'assemblage du modèle terminé, il faut minimiser l'énergie de la protéine ⇒ stéréochimie canonique

PG = Charmm, X-Plor, Amber, Gromos

Energies de liaison :

$$\sum_{liaisons} K_r (r - r_{eq})^2$$



Angles de liaison :

$$\sum_{angles\_valences} K_\theta (\theta - \theta_{eq})^2$$

Energies de torsion :

$$\sum_{angles\_dièdres} \sum_{n=1}^{3} \frac{K_n}{2} (1 + \cos(n\phi - \varphi_n))$$

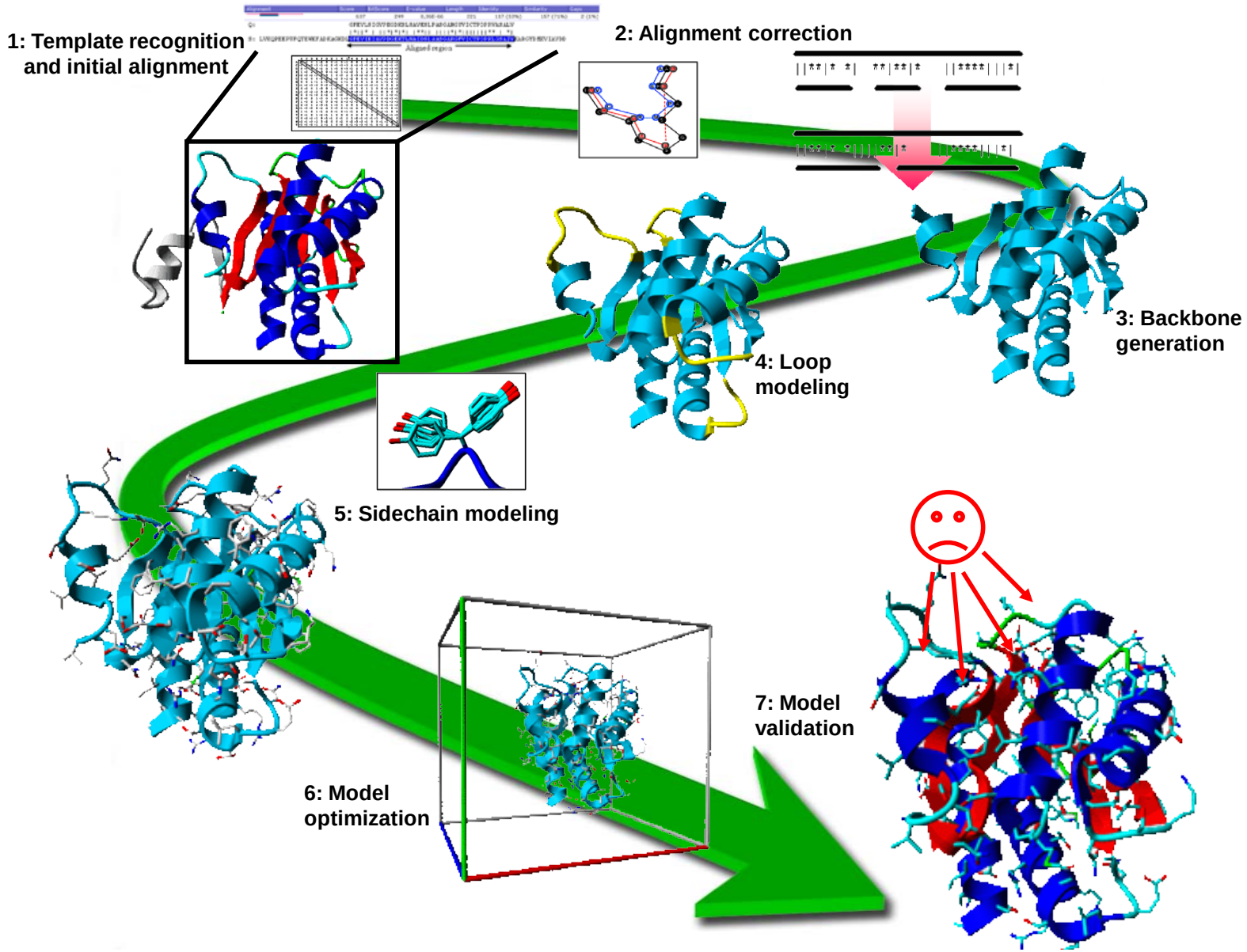
Liaisons de van der Waals :

$$\sum_{atomes\_non\_liés} \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}}$$

Interaction électrostatiques :

$$\sum_{atomes} \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r r_{ij}}$$

Évaluation du modèle



Hanka Venselaar, `http://swift.cmbi.ru.nl/teach/B1SEM/B1S_homology_modelling.ppt`

Outils pour vérifier un modèle

- Précision/Exactitude locale: ANOLEA, QMEAN, Gromos
- Qualité globale: DFire
- Stéréochimie: Whatcheck, Procheck
- Structure (super-)secondaires: DSSP, Promotif
- ModEval, ModFOLD
- Meilleurs à CASP 10: ProQ2clust2 et IntFOLD2 (ModFOLD)
- Meilleurs à CASP 12: MESHI, ProQ3, SVMQA

Meilleures méthodes de validation

- IntFOLD: `http://www.reading.ac.uk/bioinf/IntFOLD/`
- ProQ2: `http://www.bioinfo.ifm.liu.se/ProQ2/`
- ProQ3: `http://proq3.bioinfo.se/`
- SVMQA: `http://lee.kias.re.kr/SVMQA/SVMQA_eval.tar.gz`

4 Programmes / Serveurs

- Banques de données / outils
- Programmes / Serveurs - Introduction
- CASP
- SWISS-MODEL
- Logiciels graphiques locales

Banques de données / outils

Banques de données

Sequence databases	
Uniprot (141)	http://www.uniprot.org
NCBI (142)	http://www.ncbi.nlm.nih.gov/
Structure databases	
PDB (2)	http://www.pdb.org
Protein structure classifications	
SCOP (10)	http://scop.mrc-lmb.cam.ac.uk/scop/
CATH (12)	http://www.cathdb.info/
SISYPHUS (28)	http://sisyphus.mrc-cpe.cam.ac.uk/
3D complex (27)	http://www.3Dcomplex.org
Structural neighbourhoods	
MMDB (142)	http://www.ncbi.nlm.nih.gov/sites/entrez?db=structure
FSN (137)	http://fatcat.burnham.org/fatcat-cgi/cgi/FSN/fsn.pl
Dali DB (135, 143)	http://ekhidna.biocenter.helsinki.fi/dali/start
COPS (136)	http://cops.services.came.sbg.ac.at/

Andreeva, "Homology Modeling", ch. 1, Methods in Mol. Biol.(2012)

Banques de données

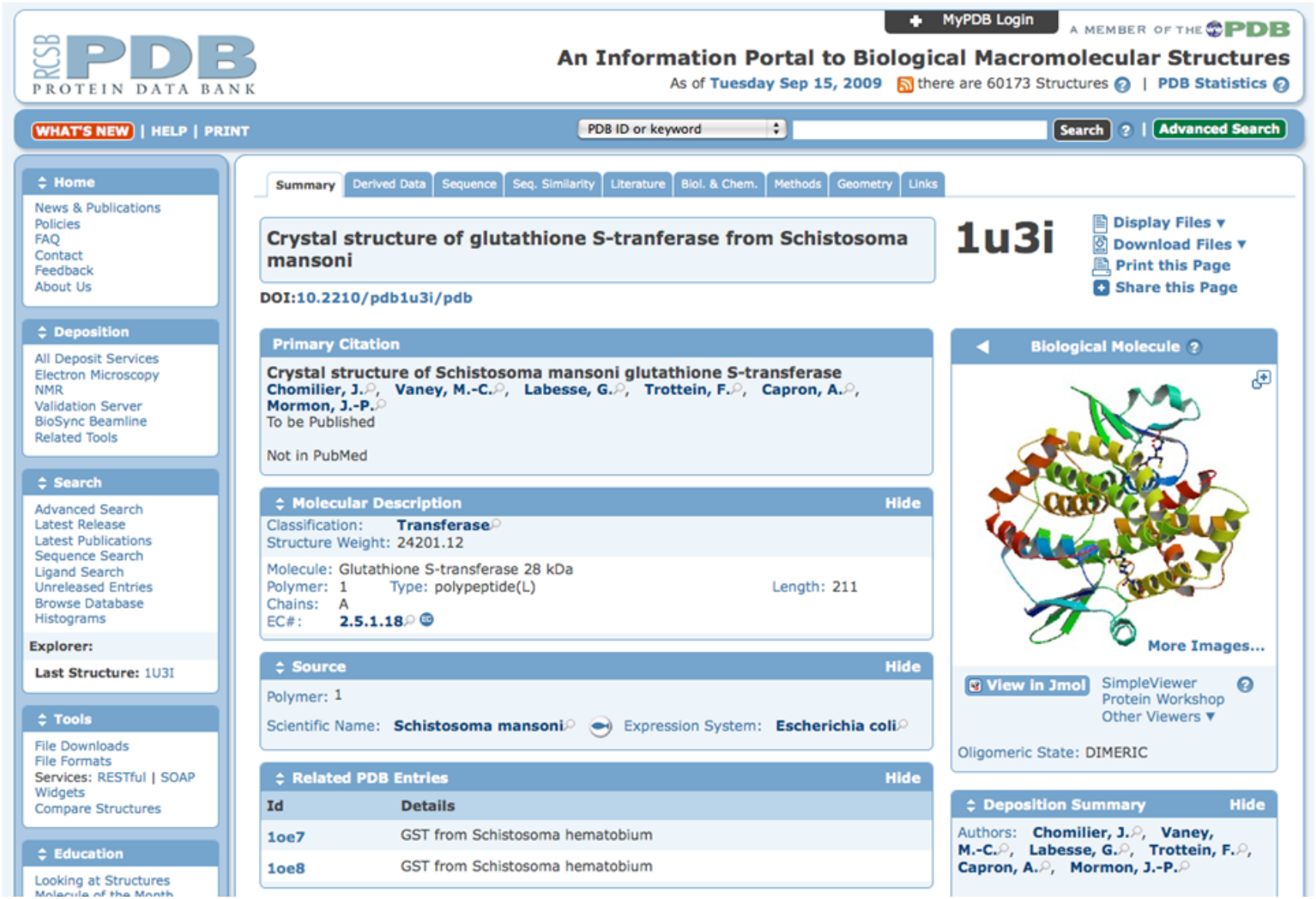
Other resources	
DisProt (84)	http://www.disprot.org/
PROSITE (26)	http://www.expasy.org/prosite
Consurf (140)	http://consurf.tau.ac.il/
Database of membrane proteins (152)	http://blanco.biomol.uci.edu/Membrane_Proteins_xtal.html
Pratt (38)	http://www.ebi.ac.uk/Tools/pratt/index.html
Jalvew (139)	http://www.jalview.org/

Andreeva, "Homology Modeling", ch. 1, Methods in Mol. Biol.(2012)

PDB, PDBsum

- PDB `http://www.rcsb.org`
- PDBsum `http://www.ebi.ac.uk/pdbsum/`

PDB


MyPDB Login
A MEMBER OF THE PDB
RCSB PDB
PROTEIN DATA BANK
An Information Portal to Biological Macromolecular Structures
As of Tuesday Sep 15, 2009 there are 60173 Structures | PDB Statistics
WHAT'S NEW | HELP | PRINT
PDB ID or keyword
Search
Advanced Search
Home
News & Publications
Policies
FAQ
Contact
Feedback
About Us
Deposition
All Deposit Services
Electron Microscopy
NMR
Validation Server
BioSync Beamline
Related Tools
Search
Advanced Search
Latest Release
Latest Publications
Sequence Search
Ligand Search
Unreleased Entries
Browse Database
Histograms
Explorer:
Last Structure: 1U3I
Tools
File Downloads
File Formats
Services: RESTful | SOAP
Widgets
Compare Structures
Education
Looking at Structures
Summary
Derived Data
Sequence
Seq. Similarity
Literature
Biol. & Chem.
Methods
Geometry
Links
Crystal structure of glutathione S-tranferase from Schistosoma mansoni
1u3i
Display Files
Download Files
Print this Page
Share this Page
DOI:10.2210/pdb1u3i/pdb
Primary Citation
Crystal structure of Schistosoma mansoni glutathione S-transferase
Chomilier, J., Vaney, M.-C., Labesse, G., Trottein, F., Capron, A., Mormon, J.-P.
To be Published
Not in PubMed
Molecular Description
Hide
Classification: Transferase
Structure Weight: 24201.12
Molecule: Glutathione S-transferase 28 kDa
Polymer: 1 Type: polypeptide(L) Length: 211
Chains: A
EC#: 2.5.1.18
Source
Hide
Polymer: 1
Scientific Name: Schistosoma mansoni Expression System: Escherichia coli
Related PDB Entries
Hide
Id
Details
1oe7
GST from Schistosoma hematobium
1oe8
GST from Schistosoma hematobium
Biological Molecule
More Images...
View in Jmol
SimpleViewer
Protein Workshop
Other Viewers
Oligomeric State: DIMERIC
Deposition Summary
Hide
Authors: Chomilier, J., Vaney, M.-C., Labesse, G., Trottein, F., Capron, A., Mormon, J.-P.

Fichier PDB

Coordonnées atomiques disponibles dans une banque de structures, la Protein Data Bank http://www.rcsb.org/pdb/

```
HEADER   DNA-BINDING PROTEIN                 20-MAY-94  1ENH     1ENH  2
COMPND   ENGRAILED HOMEODOMAIN                                    1ENH  3
SOURCE   (DROSOPHILA MELANOGASTER) RECOMBINANT FORM EXPRESSED IN  1ENH  4
SOURCE  2 (ESCHERICHIA COLI)                                      1ENH  5
AUTHOR   N.D.CLARKE,C.R.KISSINGER,J.DESJARLAIS,G.L.GILLILAND,C.O.PABO 1ENH  6
REVDAT  1   31-AUG-94 1ENH    0                                    1ENH  7
SCALE3     0.000000 0.000000 0.008466      0.00000               1ENH 61
ATOM     1  N   ARG     3      2.937  44.573  53.291  1.00 62.68    1ENH 62
ATOM     2  CA  ARG     3      3.220  44.968  51.871  1.00 61.88    1ENH 63
ATOM     3  C   ARG     3      1.922  45.475  51.229  1.00 62.67    1ENH 64
ATOM     4  O   ARG     3      0.984  44.702  51.017  1.00 65.49    1ENH 65
ATOM     5  CB  ARG     3      3.758  43.763  51.101  1.00 58.73    1ENH 66
ATOM     6  CG  ARG     3      3.642  43.884  49.610  1.00 57.06    1ENH 67
ATOM     7  CD  ARG     3      3.776  42.528  48.965  1.00 54.58    1ENH 68
ATOM     8  NE  ARG     3      5.083  42.365  48.340  1.00 56.07    1ENH 69
ATOM     9  CZ  ARG     3      6.183  41.961  48.980  1.00 57.06    1ENH 70
ATOM    10  NH1 ARG     3      6.141  41.670  50.274  1.00 57.63    1ENH 71
ATOM    11  NH2 ARG     3      7.335  41.841  48.325  1.00 57.77    1ENH 72
```

Natom Nrésidu X Y Z Q B

60

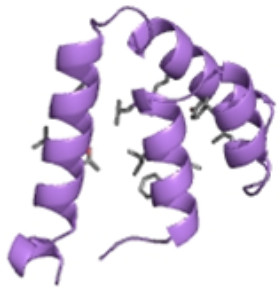
PDBSUM

Go to PDB code: 1utg go

Top page | Protein | Clefts | Links

Steroid binding — PDB id: 1utg

http://www.ebi.ac.uk/pdbsum/



Asymmetric unit

Jmol | Strap

Biological unit*, dimer
(*as deduced by PQS)

Contents

- Description
  - Header details
  - Header records
  - References
  - PROCHECK
- Protein chain
  - 70 a.a.
- Waters ×83

Tools

- Image Generation
- AstexViewer™@MSD-EBI
- Run PROCHECK
- Clefts Calculation

PDB id: 1utg
Name: Steroid binding
Title: Refinement of the c2221 crystal form of oxidized uteroglobin at 1.34 angstroms resolution
Structure: Uteroglobin. Chain: a. Engineered: yes
Source: Oryctolagus cuniculus
Biological unit: Dimer (from PQS)
UniProt: P02779 (UTER_RABIT) [Pfam]

Seq: Uteroglobin 91 a.a.
Struc: 70 a.a.

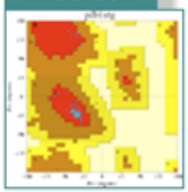
Key: Family | PfamA domain | Secondary structure

Function: (see GO annotation below)
Resolution: 1.34Å
R-factor: 0.230
Authors: I.Morize,E.Surcouf,M.C.Vaney,M.Buehner,J.P.Mornon
Key ref: I.Morize et al. (1987). Refinement of the C222(1) crystal form of oxidized uteroglobin at 1.34 A resolution.. *J Mol Biol*, **194**, 725-739. [PubMed id: 3656405] [DOI: 10.1016/0022-

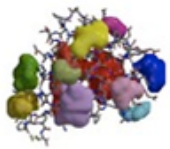
Quick links

- RCSB
- MSD
- SRS
- MMDB
- JenaLib
- OCA
- Proteopedia
- CATH
- SCOP
- FSSP
- HSSP
- PQS
- ProSAT
- Whatcheck

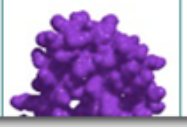
Procheck



Clefts



Surface



Outils

Tools for analysis	
Tools for sequence comparison and similarity searches	
BLAST & PSIBLAST (85)	http://www.ncbi.nlm.nih.gov/blast
FASTA3 (144)	http://www.ebi.ac.uk/Tools/fasta33
HMMER (86)	http://selab.janelia.org/
Tools for structure comparison and similarity searches	
Dali (143)	http://ekhidna.biocenter.helsinki.fi/dali_server/
VAST (145)	http://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html
SSAP (146)	http://www.cathdb.info
FATCAT (147)	http://fatcat.burnham.org/
CE (148)	http://cl.sdsc.edu/
Mammoth (149)	http://ub.cbm.uam.es/mammoth/mult/
Topmatch (150)	http://topmatch.services.came.sbg.ac.at/TopMatchFlex.php
TM-align (151)	http://zhanglab.ccmb.med.umich.edu/TM-align/

Andreeva, "Homology Modeling", ch. 1, Methods in Mol. Biol.(2012)

Comparaison de structure 3D

DALI server

`http://ekhidna.biocenter.helsinki.fi/dali_server/start`

DALI banque

`http://ekhidna.biocenter.helsinki.fi/dali/start`

CE

`http://source.rcsb.org/jfatcatserver/ceHome.jsp`

3D-Blast

`http://threedblast.loria.fr/`

YAKUSA

`http://bioserv.rpbs.jussieu.fr/Yakusa/index.html`

Programmes / Serveurs - Introduction

Table 1. Commonly Used Tools and Services for Protein Structure Modeling and Prediction

Tool or Service	Web Site
Protein Model Portal	http://www.proteinmodelportal.org (Arnold et al., 2009; Haas et al., 2013)
Model Archive	http://modelarchive.org
HHpred	http://toolkit.tuebingen.mpg.de/hhpred (Hildebrand et al., 2009)
IMP	http://www.salilab.org/imp (Russel et al., 2012; Yang et al., 2012)
IntFOLD	http://www.reading.ac.uk/bioinf/IntFOLD/ (Roche et al., 2011)
I-Tasser	http://zhanglab.ccmb.med.umich.edu/I-TASSER/ (Zhang, 2013)
ModBase	http://salilab.org/modbase/ (Pieper et al., 2011)
Modeler/ModWeb	http://salilab.org/modeller/ (Pieper et al., 2011; Yang et al., 2012)
Pcons.net	http://pcons.net/ (Larsson et al., 2011)
PHYRE2	http://www.sbg.bio.ic.ac.uk/phyre2/ (Kelley and Sternberg, 2009)
Robetta	http://robetta.bakerlab.org/ (Raman et al., 2009)
Rosetta	https://www.rosettacommons.org (Das and Baker, 2008)
SWISS-MODEL Repository	http://swissmodel.expasy.org/repository (Kiefer et al., 2009)
SWISS-MODEL Workspace	http://swissmodel.expasy.org/workspace/ (Arnold et al., 2006; Bordoli and Schwede, 2012)

Torsten Schwede (Sept. 2013). In: *Structure* 21.9

CASP

CASP - Critical Assessment of protein Structure Prediction

- `http://predictioncenter.org/`
- tous les deux ans:
  CASP 10: 2012, CASP 11: 2014, CASP 12: 2016, CASP 13: 2018
- et depuis 2011: en continue => CASP ROLL

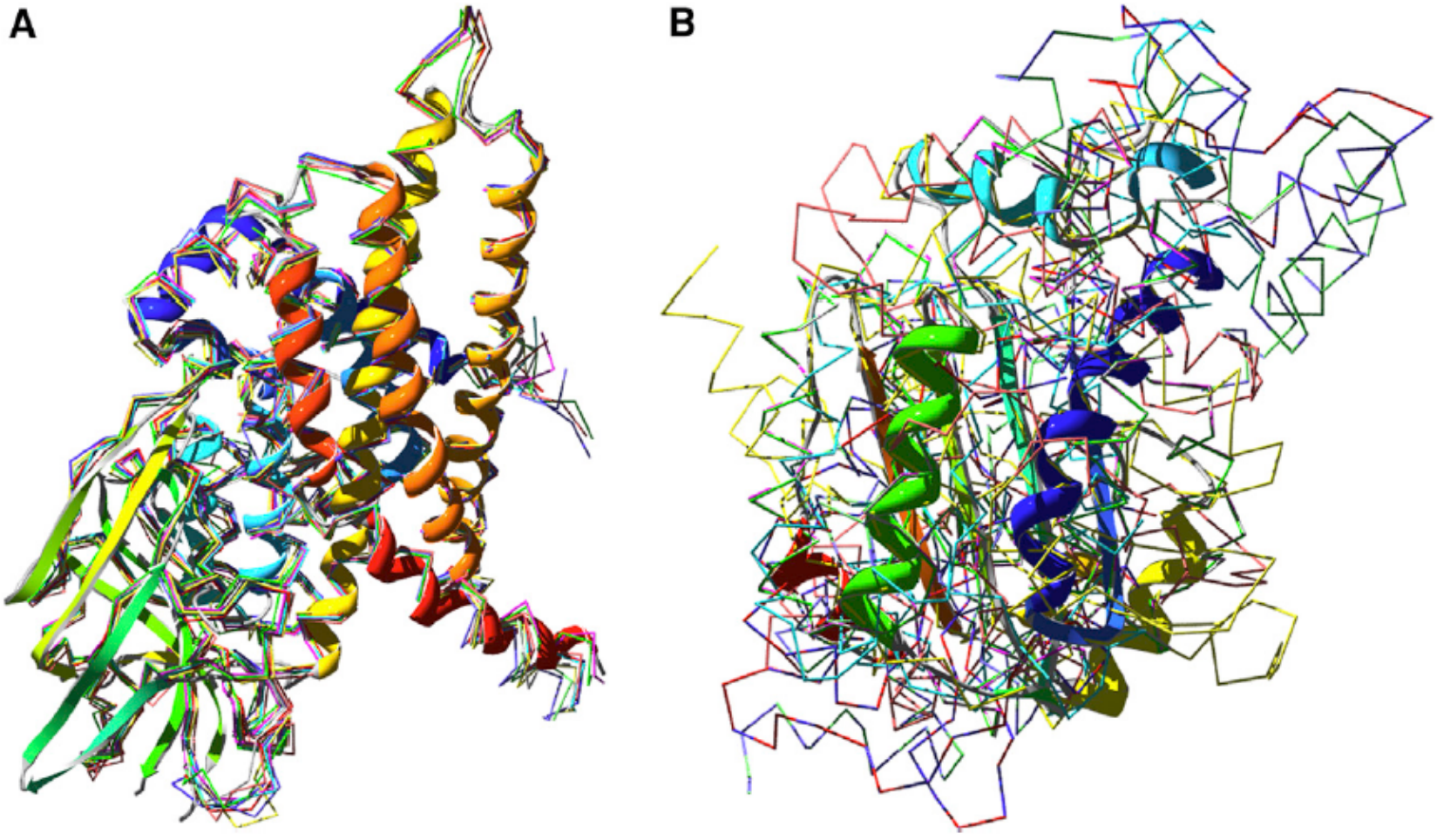
CASP - Modeling categories

1. Template-based modeling (TBM)
   1. comparative modeling
   2. fold recognition
2. Free modeling (FM)
   1. Knowledge-based *de novo* modeling
   2. *ab initio* modeling from first principles
3. Refinement

CASP - Other modeling categories

1. Contact-assisted structure modeling
2. Chemical shifts guided modeling of NMR structures
3. Structure modeling based on molecular replacement with ab initio models and X-ray data
4. Detecting residue-residue contacts in proteins (RR).
5. Identifying disordered regions in target proteins (DR).
6. Function prediction (prediction of binding sites) (FN).
7. Quality assessment of models in general (without knowing native structures) and the reliability of predicting certain residues in particular (QA).

CASP 10 - easy VS difficult target for TBM



Torsten Schwede (Sept. 2013). In: *Structure* 21.9

CASP - Liste des meilleurs serveurs

I-TASSER - http://zhanglab.ccmb.med.umich.edu/I-TASSER/
ROBETTA - http://robetta.bakerlab.org/
HHpred - http://toolkit.tuebingen.mpg.de/hhpred/
METATASSER - http://cssb.biology.gatech.edu/skolnick/webservice/MetaTASSER/
MULTICOM - http://casp.rnet.missouri.edu/multicom_3d.html
Pcons - http://pcons.net/
SAM-T08 - http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html
3D-Jury - http://meta.bioinfo.pl/submit_wizard.pl
THREADER - http://bioinf.cs.ucl.ac.uk/threader/
RaptorX - http://raptorx.uchicago.edu/
Autre serveurs populaires:

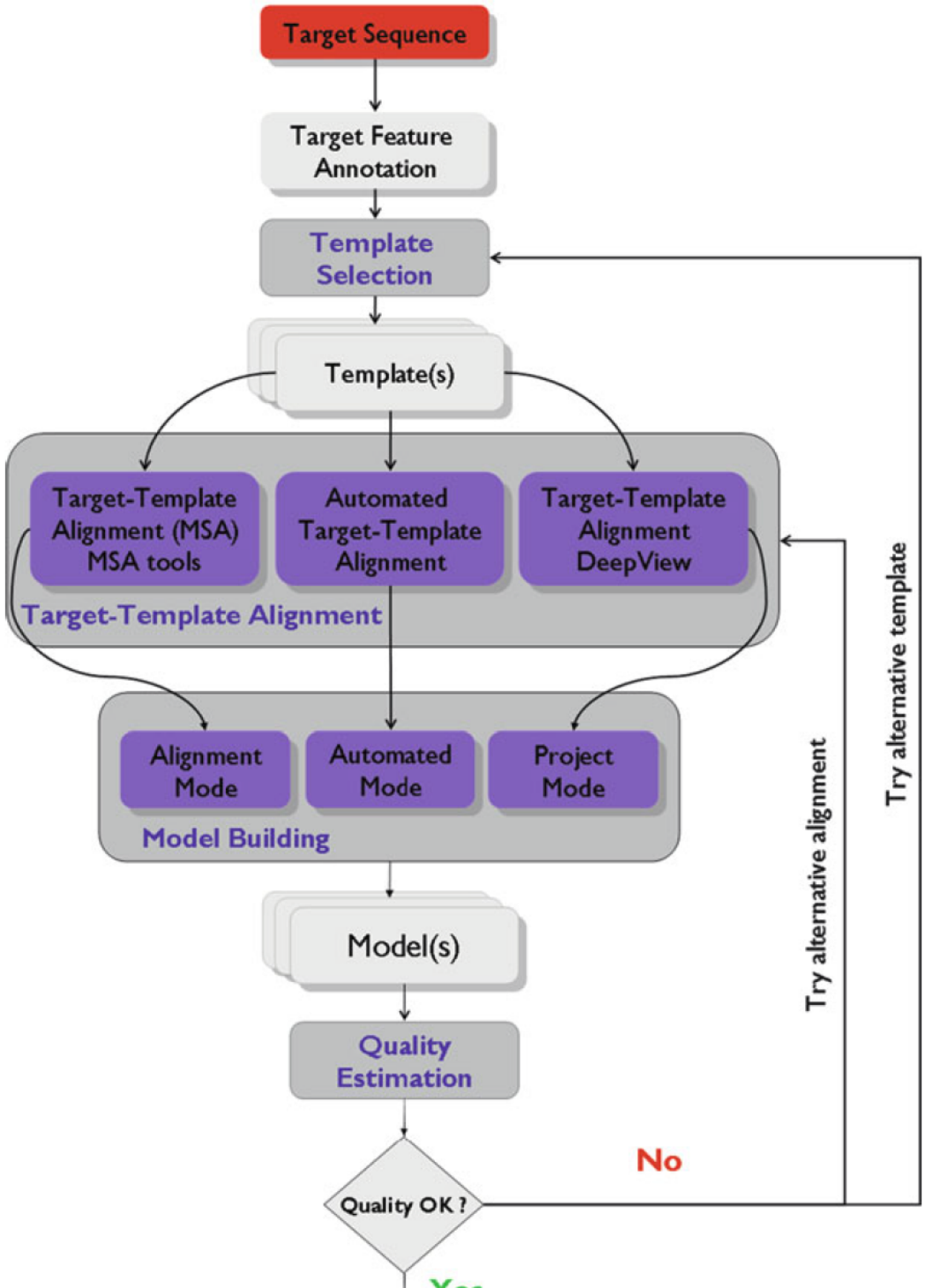
- SwissModel
- MODELLER

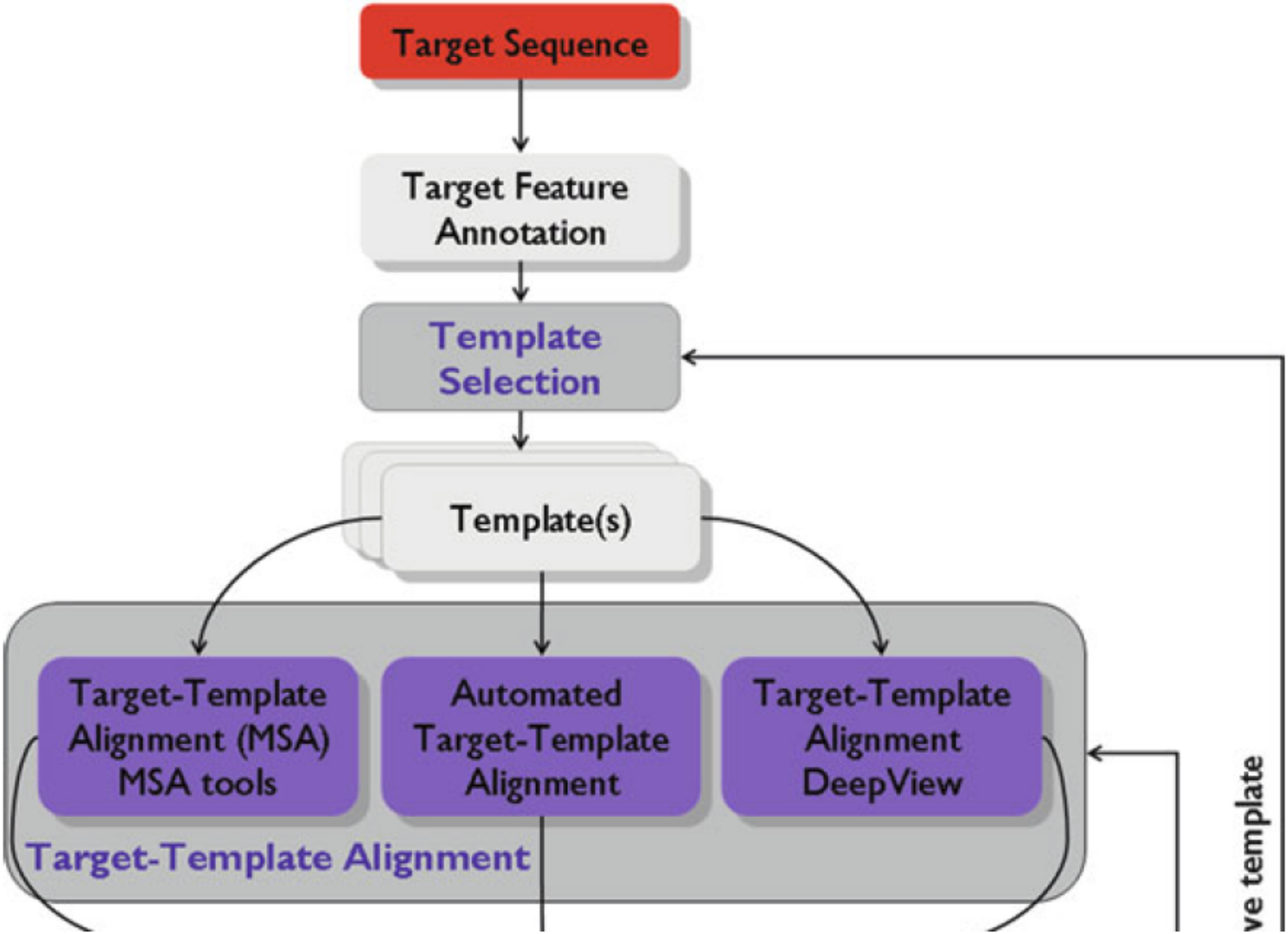
SWISS-MODEL

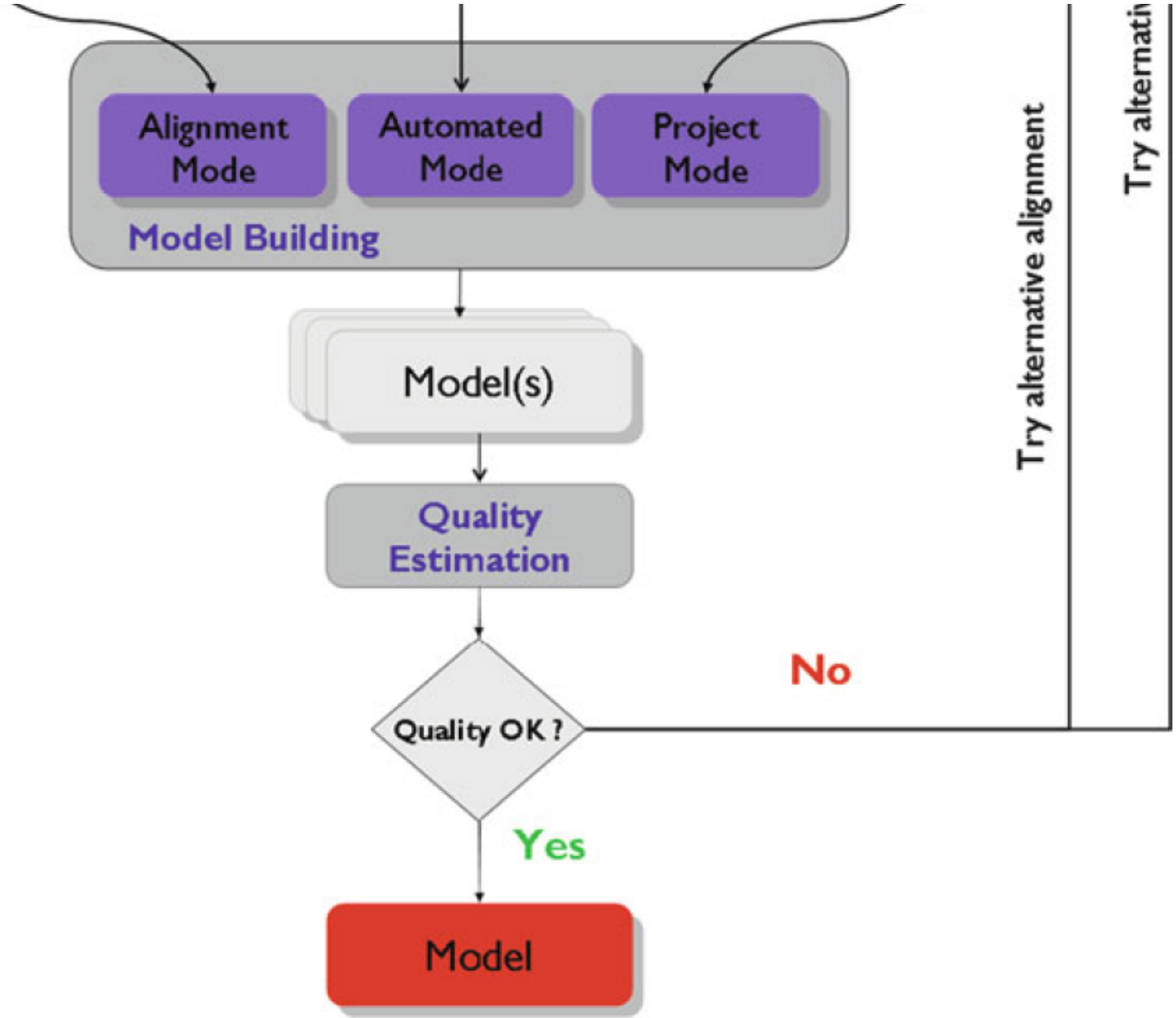
Introduction

`http://swissmodel.expasy.org/`

- Web-serveur automatisé pour la modélisation par homologie
- Existe depuis 20 ans
- Dernière version: SWISS-MODEL Workspace (2006)
- Accessible pour les non-experts en bioinfo
- Automatisation permet une meilleure reproductibilité


Target Sequence
Target Feature Annotation
Template Selection
Template(s)
Target-Template Alignment (MSA) MSA tools
Automated Target-Template Alignment
Target-Template Alignment DeepView
Target-Template Alignment
Alignment Mode
Automated Mode
Project Mode
Model Building
Model(s)
Quality Estimation
Quality OK ?
No
Try alternative alignment
Try alternative template


Target Sequence
Target Feature Annotation
Template Selection
Template(s)
Target-Template Alignment (MSA) MSA tools
Automated Target-Template Alignment
Target-Template Alignment DeepView
Target-Template Alignment
ve template


Alignment Mode
Automated Mode
Project Mode
Model Building
Model(s)
Quality Estimation
Quality OK ?
No
Yes
Model
Try alternative alignment
Try alternati

Trois modes

En fonction de la distance évolutive, on choisit un des trois modes:

1. Automated: que besoin de la séquence de la cible, le génération du modèle par homologie est entièrement automatisé. Conseillé pour les cas simples (identité de séquence >60%)
2. Alignment: MSA corrigé manuellement doit être fournit. Conseillé pour des cas non-triviaux.
3. Project: L'alignement cible-template peut-être inspecté et modifié avec DeepView (= Swiss-Pdb Viewer)

Outils

1. Annotation (séquence, structure, fonction):
   - InterProScan: Identification domaines, motifs, familles
   - PsiPred: Prédiction structure secondaire
   - DisoPred: Prédiction désordre (=> IUP)
   - MEMSAT: Prédiction ségments transmembranaires
2. Alignement:
   - BLAST, PSI-BLAST, HHsearch contre la SMTL: SWISS-MODEL Template Library
   - DeepView (=Swiss-Pdb Viewer) pour la correction manuelle
   - Fournit par l'utilisateur: MSA par outils extérieurs
3. Validation:
   - Précision/Exactitude locale: ANOLEA, QMEAN, Gromos
   - Qualité globale: DFire
   - Stéréochimie: Whatcheck, Procheck
   - Structure (super-)secondaires: DSSP, Promotif

Logiciels graphiques locales

Pour modélisation interactive

- DeepView (=Swiss-Pdb Viewer): `http://spdbv.vital-it.ch/`
- MolIDE: `http://dunbrack.fccc.edu/molide/`
- MolIDE2 alias BioAssemblyModeler: `http://dunbrack.fccc.edu/BAM/`
- PyMod: `http://schubert.bio.uniroma1.it/pymod/index.html`
- D'autres interfaces MODELLER: `http://salilab.org/modeller/wiki/Links`

5 Bibliography

Bibliography I

- Baltzis, Athanasios et al. (Sept. 2022). “Highly significant improvement of protein sequence alignments with AlphaFold2”. In: *Bioinformatics*, btac625.
- Carpentier, Mathilde and Jacques Chomilier (Oct. 2019). “Protein multiple alignments: sequence-based versus structure-based programs”. In: *Bioinformatics* 35.20, pp. 3970–3980.
- Deorowicz, Sebastian, Agnieszka Debudaj-Grabysz, and Adam Gudyś (Sept. 2016). “FAMSA: Fast and accurate multiple sequence alignment of huge protein families”. en. In: *Scientific Reports* 6.1. Number: 1 Publisher: Nature Publishing Group, p. 33964.
- Mirdita, M et al. (Sept. 2021). “Fast and sensitive taxonomic assignment to metagenomic contigs”. In: *Bioinformatics* 37.18, pp. 3029–3031.
- Mirdita, Milot, Konstantin Schütze, et al. (June 2022). “ColabFold: making protein folding accessible to all”. en. In: *Nature Methods* 19.6. Number: 6 Publisher: Nature Publishing Group, pp. 679–682.
- Mirdita, Milot, Martin Steinegger, and Johannes Söding (Aug. 2019). “MMseqs2 desktop and local web server app for fast, interactive sequence searches”. In: *Bioinformatics* 35.16, pp. 2856–2858.
- Schwede, Torsten (Sept. 2013). “Protein Modeling: What Happened to the “Protein Structure Gap”?” In: *Structure* 21.9, pp. 1531–1540.

Bibliography II


Steinegger, Martin and Johannes Söding (Nov. 2017). “MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets”. en. In: *Nature Biotechnology* 35.11. Number: 11 Publisher: Nature Publishing Group, pp. 1026–1028.


Taly, Jean-Francois et al. (Nov. 2011). “Using the T-Coffee package to build multiple sequence alignments of protein, RNA, DNA sequences and 3D structures”. en. In: *Nature Protocols* 6.11, pp. 1669–1682.


Warnow, Tandy (2021). “Revisiting Evaluation of Multiple Sequence Alignment Methods”. en. In: *Multiple Sequence Alignment: Methods and Protocols*. Ed. by Kazutaka Katoh. Methods in Molecular Biology. New York, NY: Springer US, pp. 299–317.