Meet-U

A meeting story

edition 2020

When protein genes are sequenced

It's Time To Fold!

http://www.meet-u.org/

SORBONNE UNIVERSITÉ

université PARIS DIDEROT

UNIVERSITÉ PARIS SUD

université PARIS-SACLAY

OBJECTIVES

By realizing a project from A to Z to address a challenging open question in biology

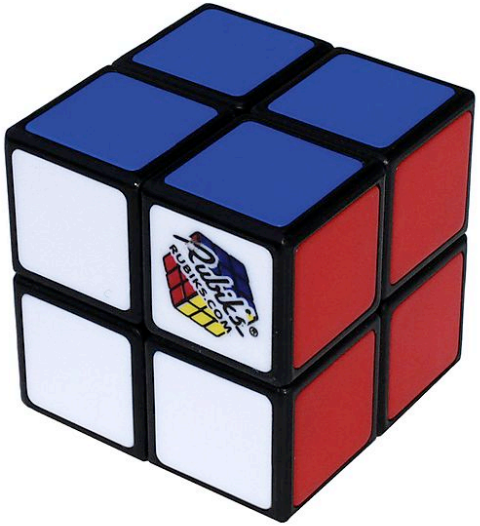
OBJECTIVES

By realizing a project from A to Z to address a challenging open question in biology

Prediction of protein 3D structures based on sequence similarity

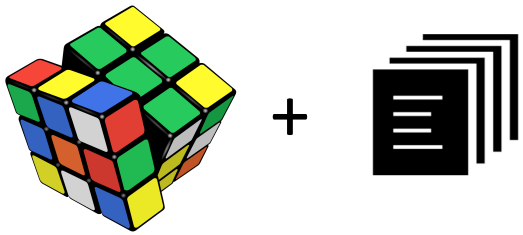
OBJECTIVES

OBJECTIVES



FRAMEWORK

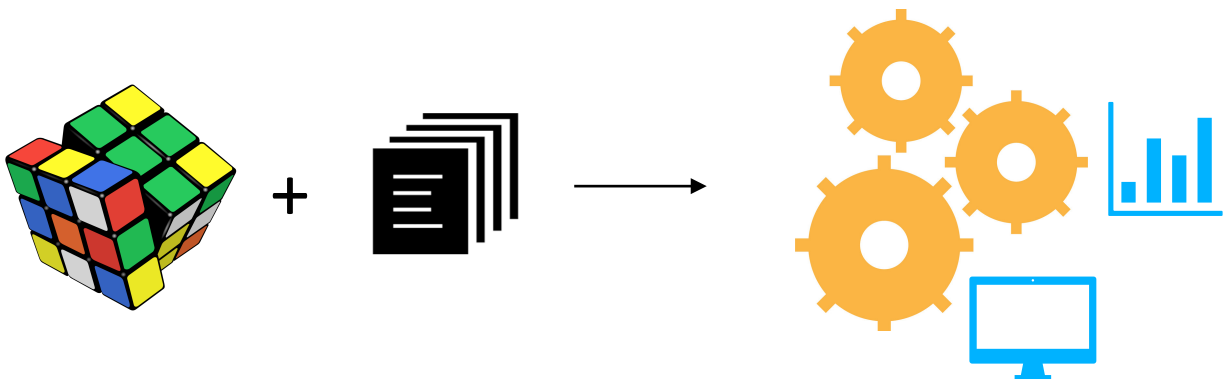
By realizing a project from A to Z to address a challenging open question in biology



Pedagogical team

FRAMEWORK

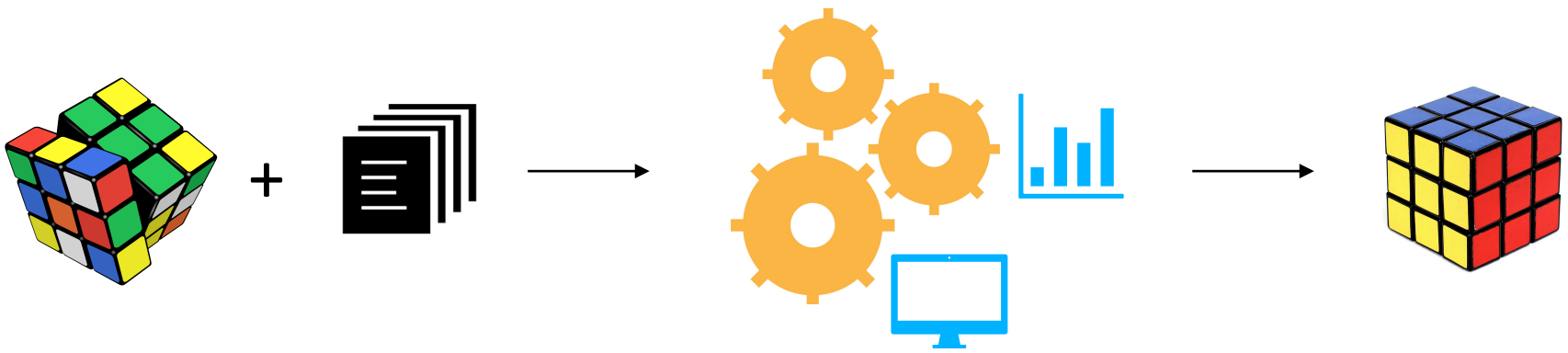
By realizing a project from A to Z to address a challenging open question in biology



Pedagogical team **Students**

FRAMEWORK

By realizing a project from A to Z to address a challenging open question in biology



Pedagogical team **Students** **Final product**

WHO?

Pedagogical Teams:

- Univ. Pierre et Marie Curie (Sorbonne Université)
  J. Bernardes, E. Laine, Y. Mohseni Behbahani, M. Muscat
- Univ. Paris Saclay
  - G. Lelandais, A. Lopes, C. Papadopoulos

Students from:

- 2 masters of bioinformatics

Jury members from France and Europe

(invitations are being sent…)

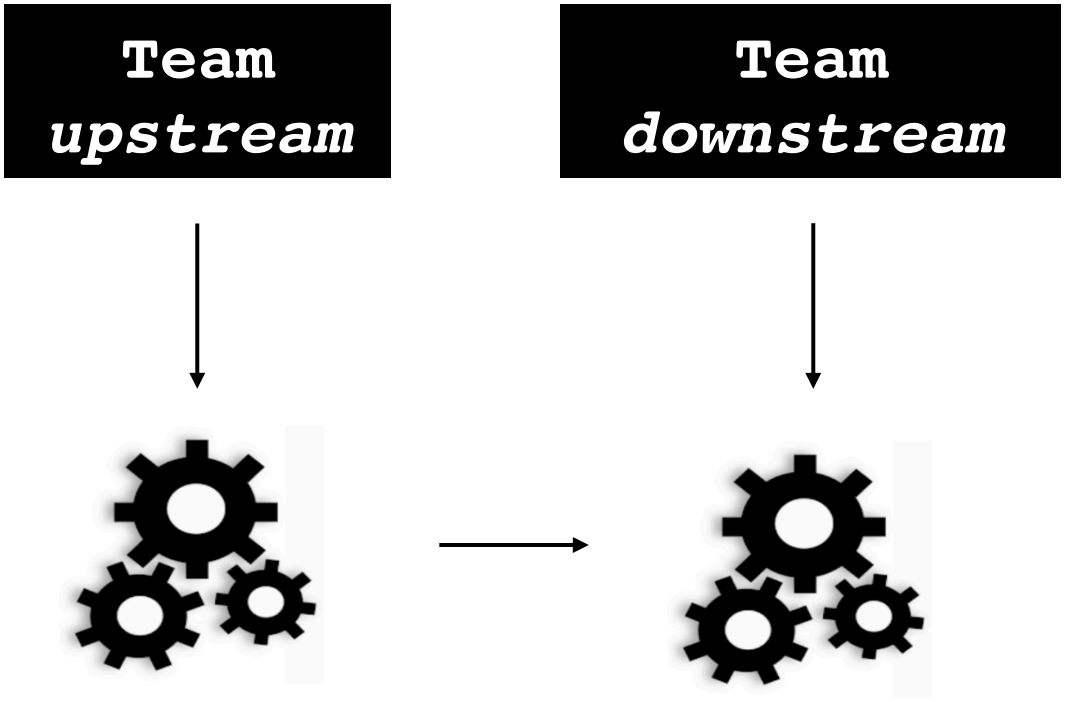
HOW?

Team *upstream*

HOW?

Team *upstream*

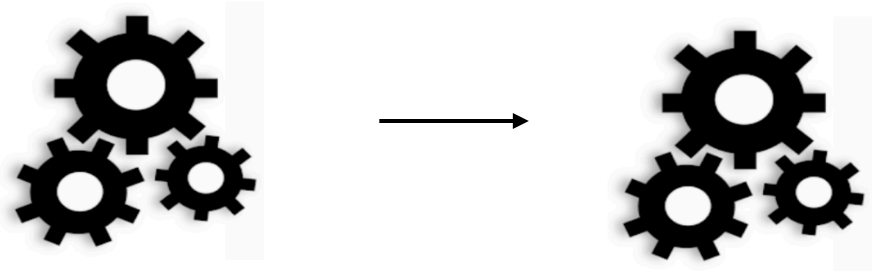
HOW?



HOW?

Team *upstream*

Team *downstream*

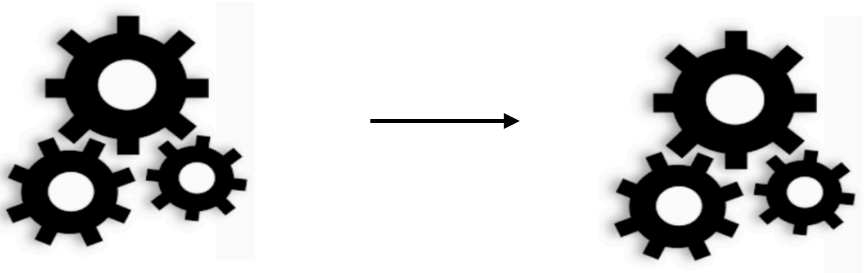


Integration half-way of the project

HOW?

Team *upstream*

Team *downstream*



Integration half-way of the project

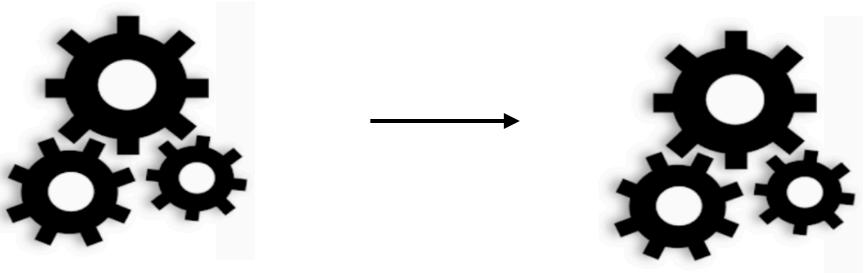
GIT/GITHUB

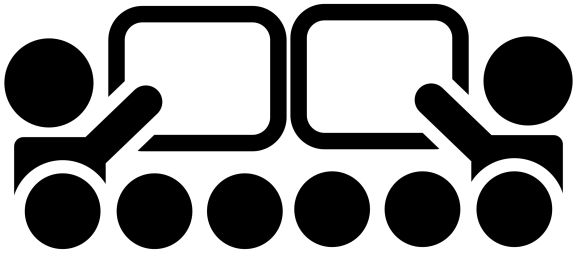
CLUSTER

HOW?

Team *upstream*

Team *downstream*





Flipped symposium

Meet-U 2020

- Prediction of the fold of a protein from its sequence

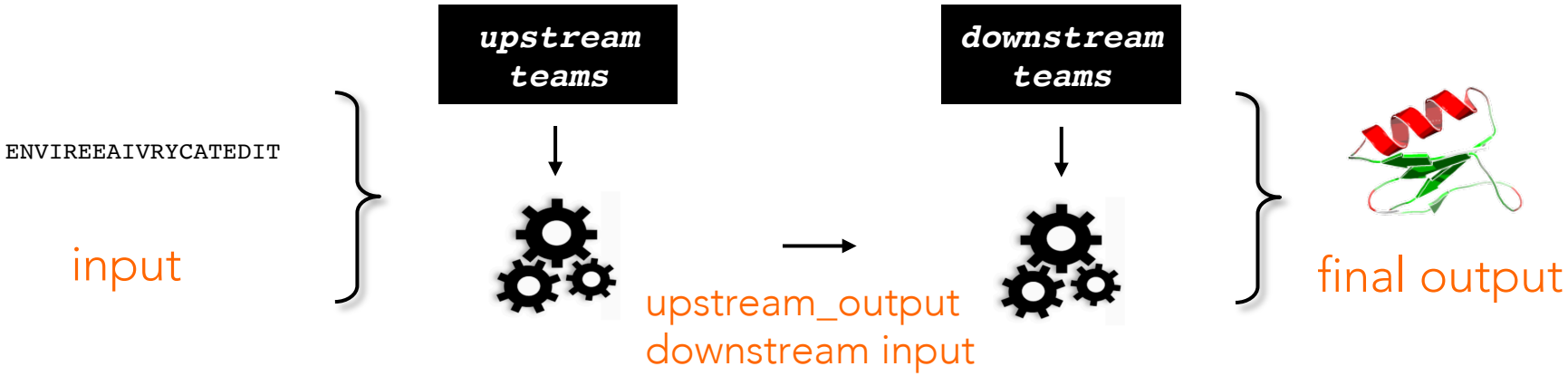
ENVIREEAIVRYCATEDIT

input

output

Meet-U 2020

3 constraints!



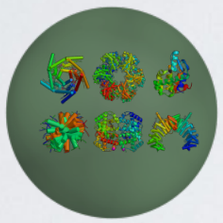
- 2 steps
  - **upstream teams:** building and querying sequence profiles against a database of protein structure families (dynamic programming)
  - **downstream teams:** assessment of profile-structure compatibility and identification of the most probable fold

~50 millions protein sequences

(UniprotKB/TrEMBL)

~100 000 protein structures

(Protein DataBank)



<<

SEQUENCE/STRUCTURE CONSERVATION

Similar sequences

```
GLLT-ESQLVIKSAWEEFN...
GVSDGDWNIVLNTSGKVEA...
GDLSLE-QKKIRSTWHQLM...
GISDGEWQLVVNASGKVES...
```

Sequence similarity : 30 %

Similar structures

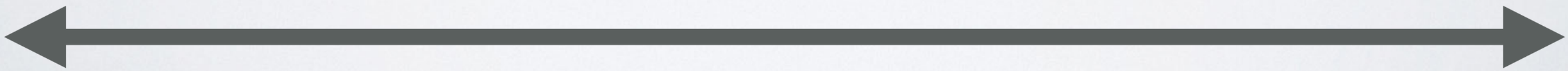
Structural similarity : ~70 %
(secondary structure conservation)

Structure is more conserved than sequence

Illergård K, Ardell DH, Elofsson A. Proteins. 2009 Nov 15;77(3):499-508

PRINCIPLE

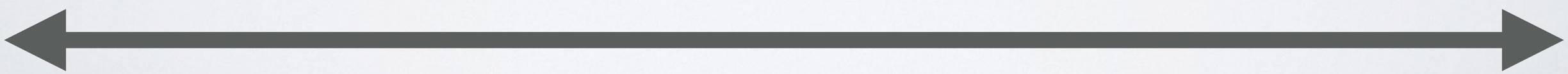
Protein A ● ● Protein B



Evolutive distance

PRINCIPLE

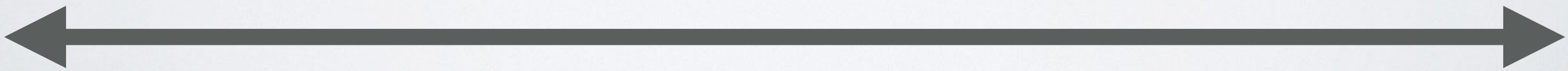
Protein A  Protein B



Evolutive distance

PRINCIPLE

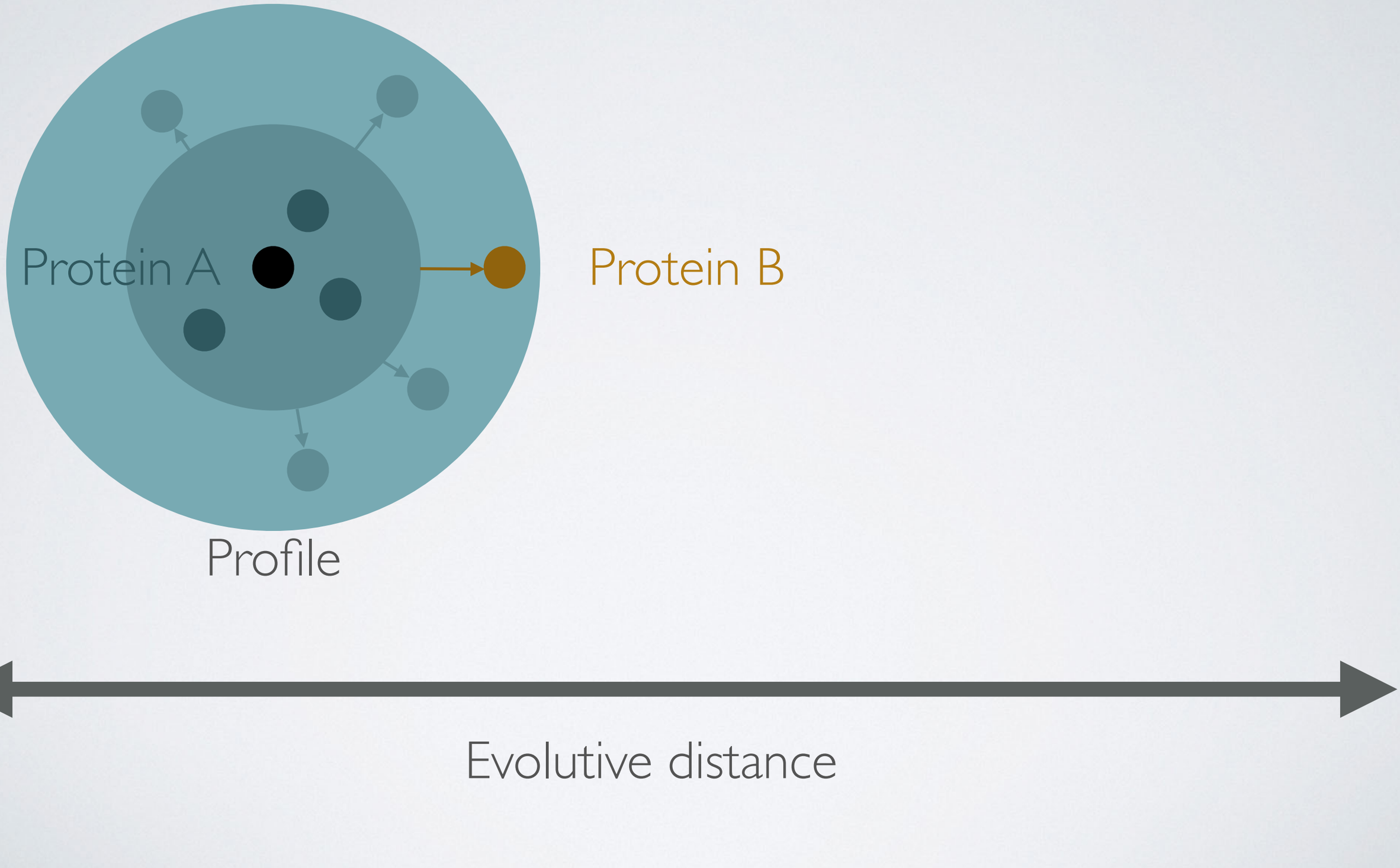
Protein A ● ● Protein B



Evolutive distance

PRINCIPLE

PRINCIPLE



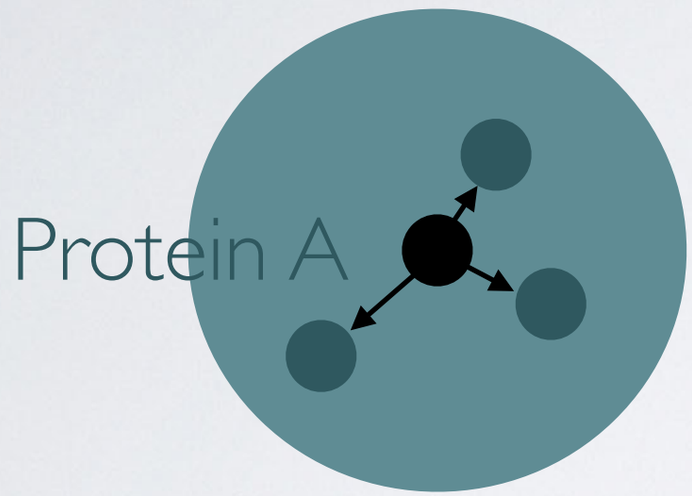
PRINCIPLE

Protein A

Protein C

Evolutive distance

PRINCIPLE

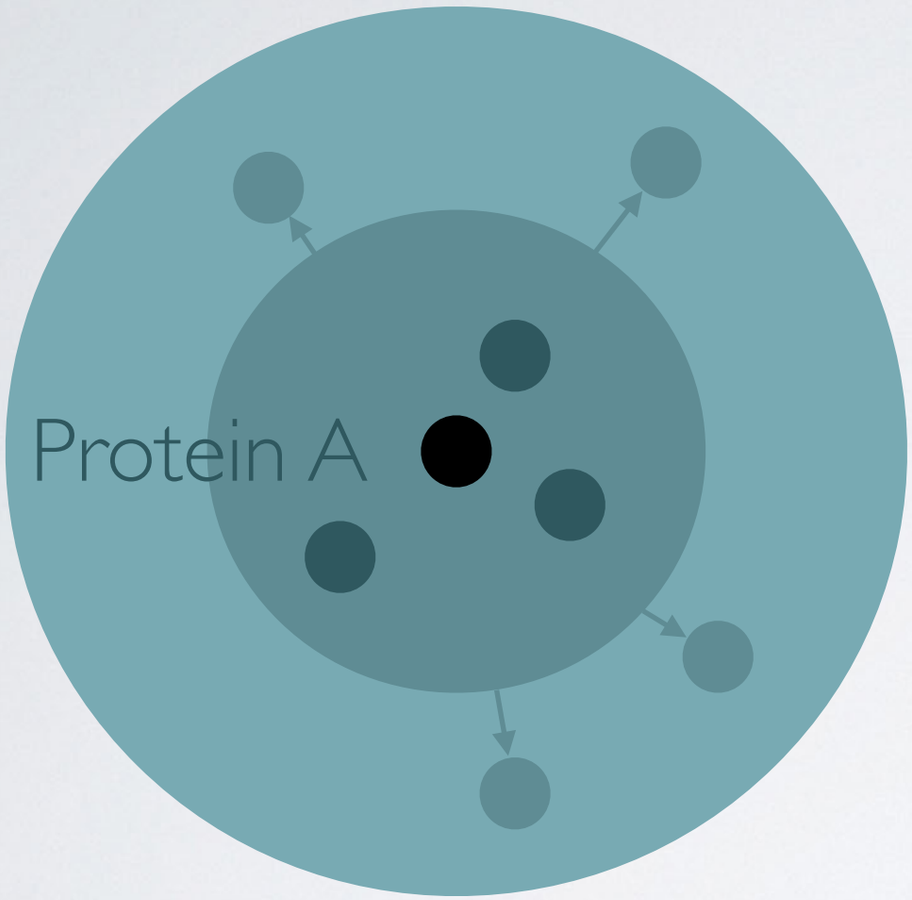


Protein C

Profile

Evolutive distance

PRINCIPLE

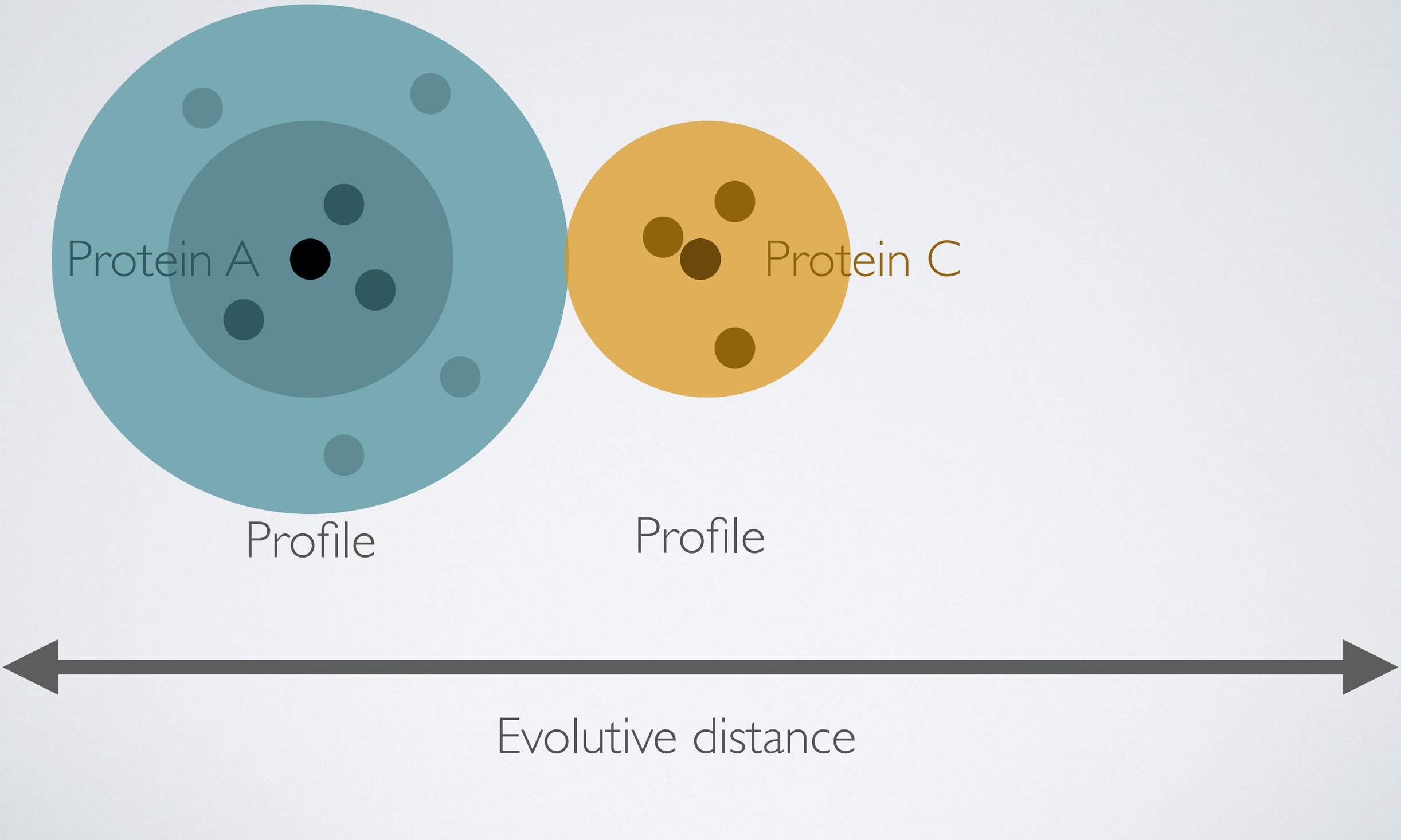


Profile

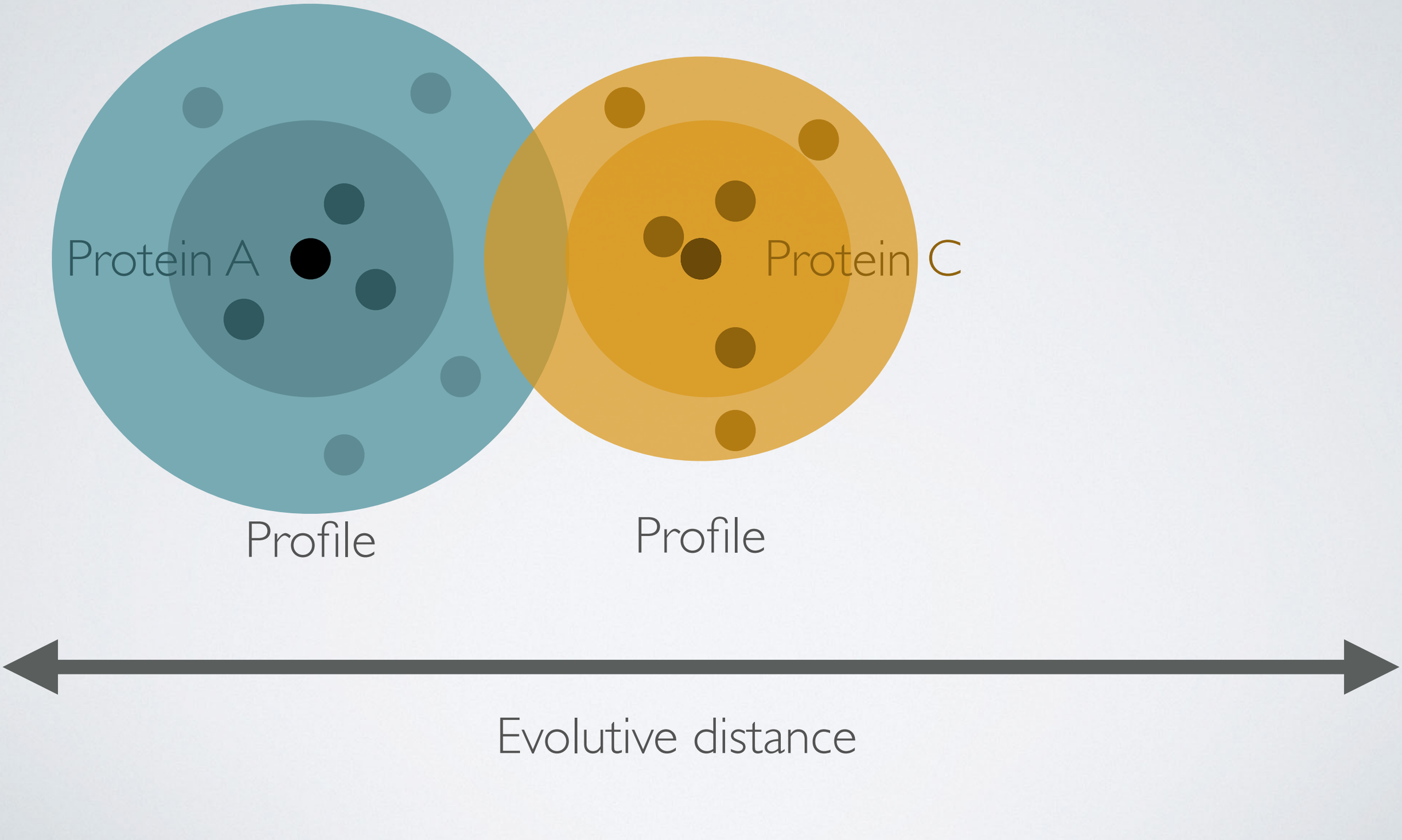
Protein C

Evolutive distance

PRINCIPLE



PRINCIPLE



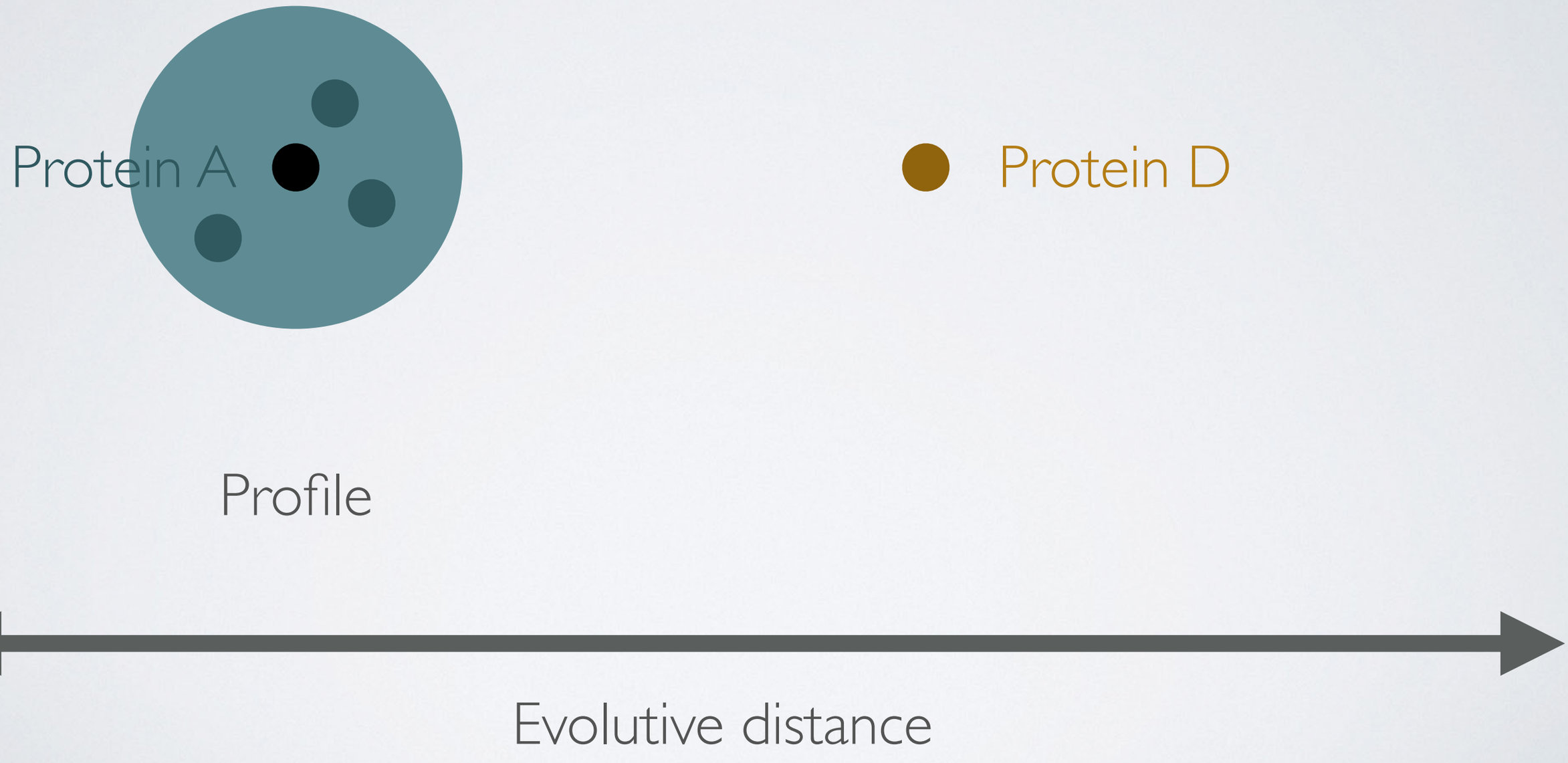
PRINCIPLE

Protein A

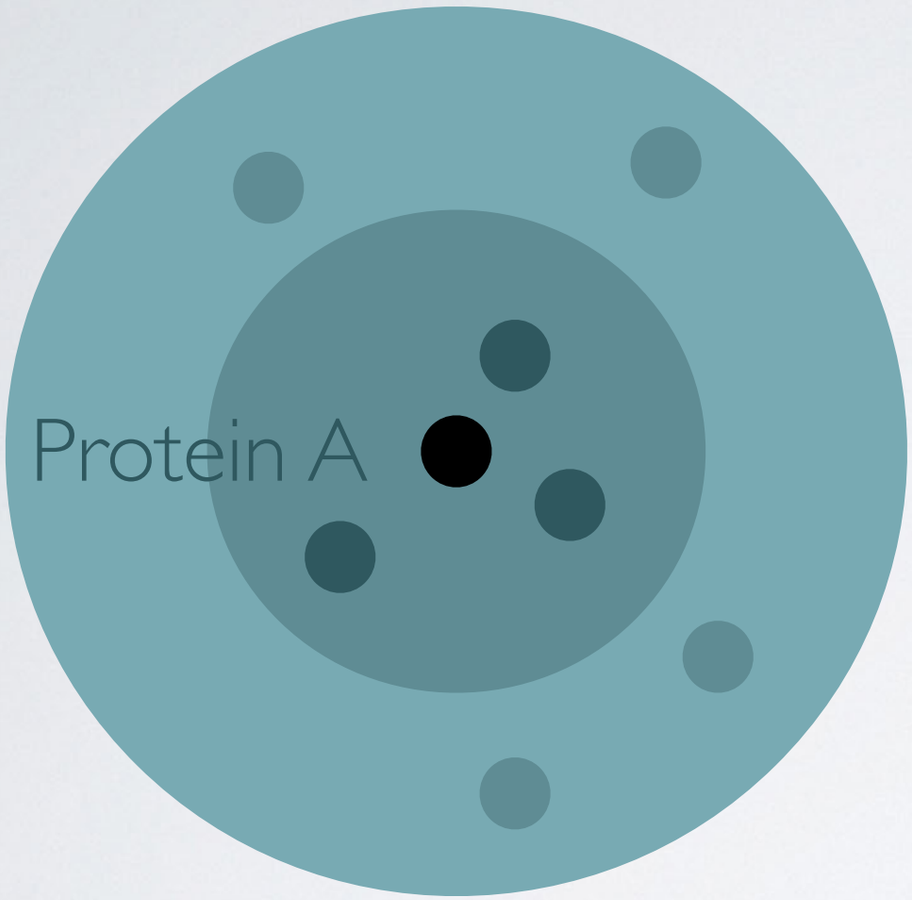
Protein D

Evolutive distance

PRINCIPLE



PRINCIPLE

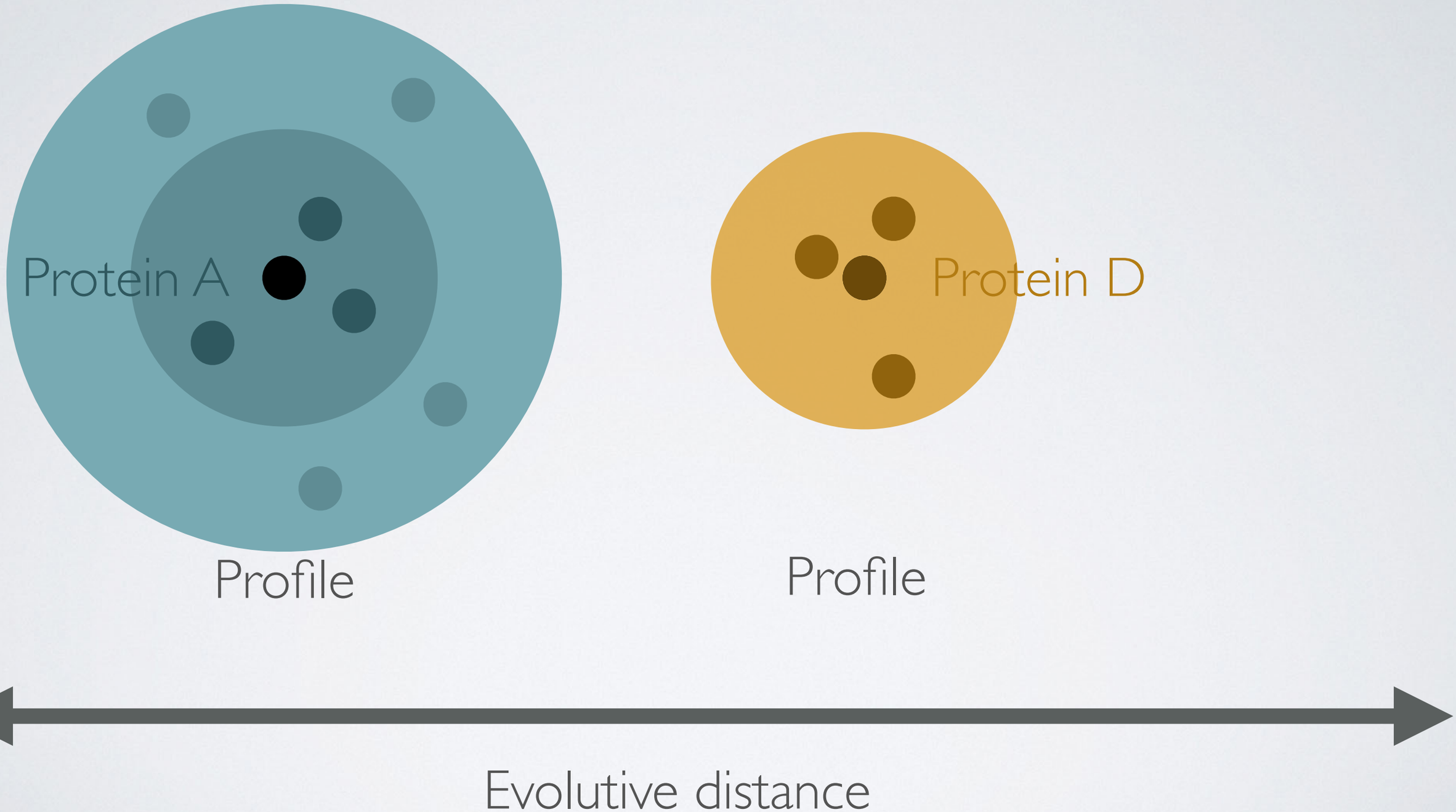


Profile

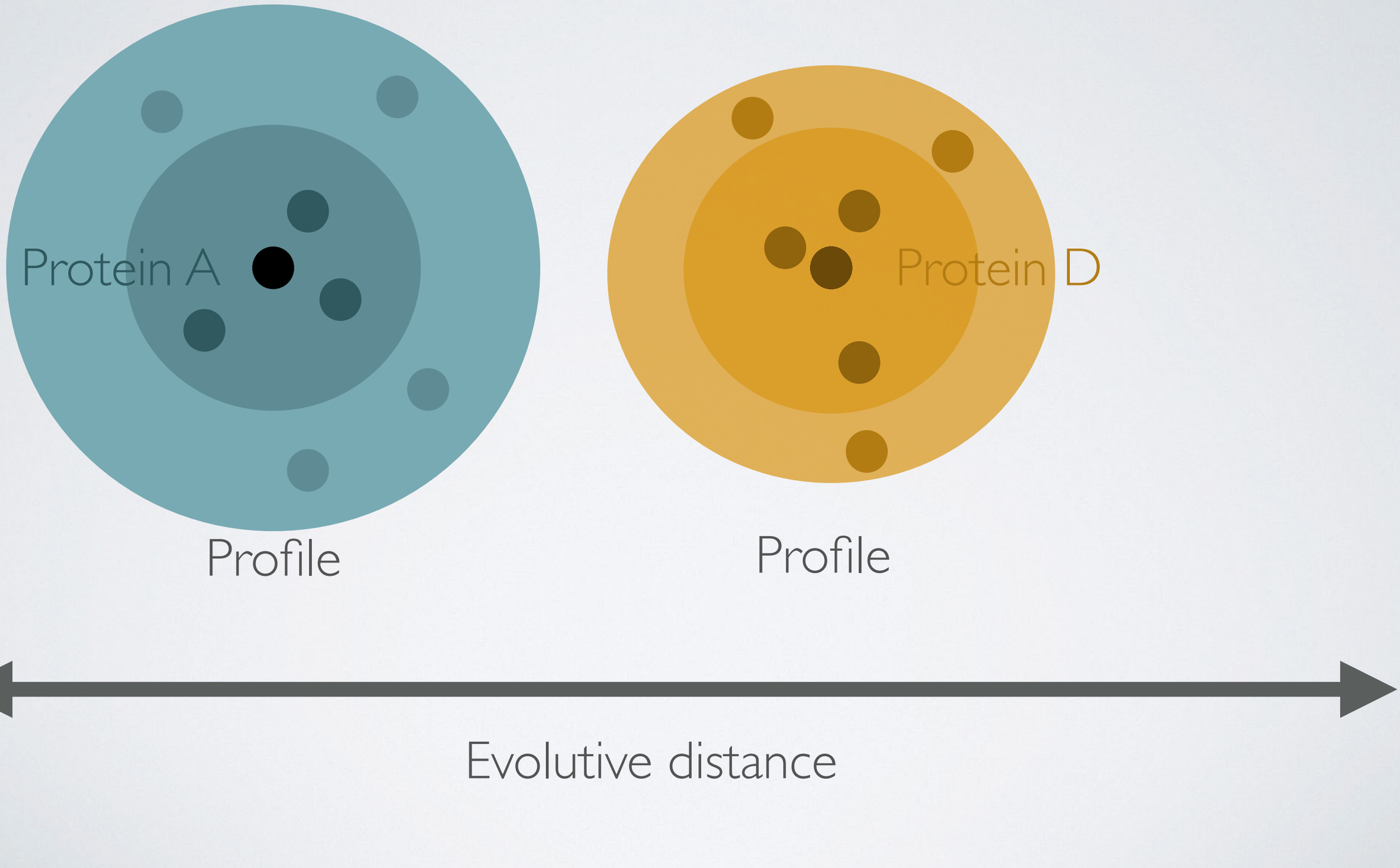
Protein D

Evolutive distance

PRINCIPLE



PRINCIPLE



HISTORY OF FOLD RECOGNITION

Query Sequence

RHPVVMGNWKLNGSKEMVVDLLNGLNAELEGVTGVDV

Template Sequences

KTVRWMLKNADSQE

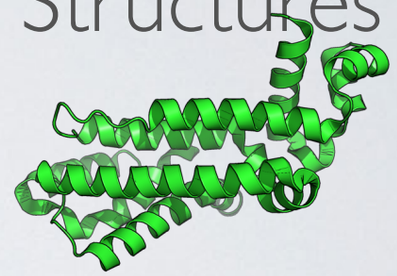
ERHKWMLKNADSQ

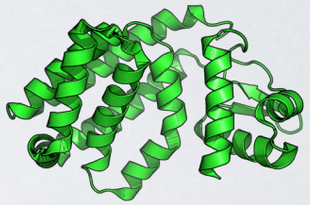
LHHKWMLKNADSQL

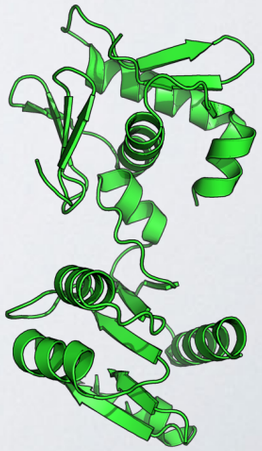
MLHKWMLKNADSQL

GFSKWMLKNADSQL

Template Structures







HISTORY OF FOLD RECOGNITION

Query Sequence

RHPVVMGNWKLNGSKEMVVDLLNGLNAELEGVTGVDV

Search

Template Sequences

KTVRWMLKNADSQE

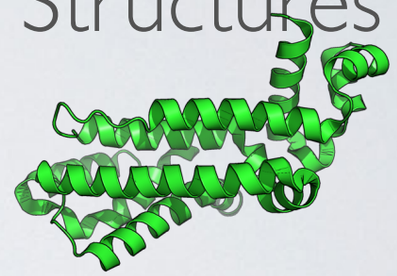
ERHKWMLKNADSQ

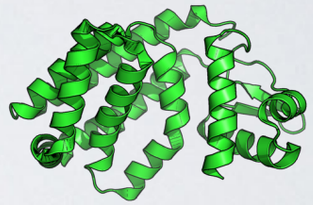
LHHKWMLKNADSQL

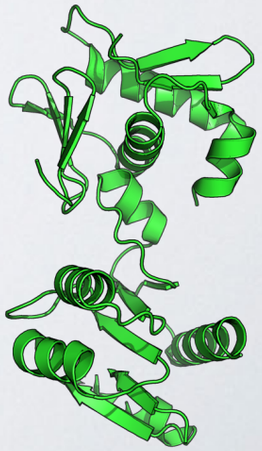
MLHKWMLKNADSQL

GFSKWMLKNADSQL

Template Structures







Sequence-Sequence

SSEARCH (Smith & Waterman)

BLAST

HISTORY OF FOLD RECOGNITION

Query Sequence

RHPVVMGNWKLNGSKEMVVDLLNGLNAELEGVTGVDV

Search

Template Sequences

KTVRWMLKNADSQE

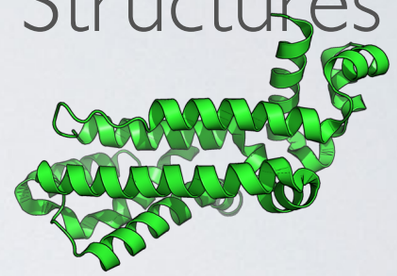
ERHKWMLKNADSQ

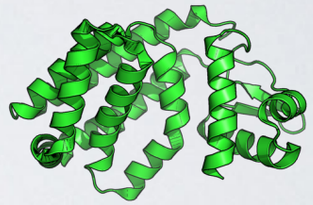
LHHKWMLKNADSQL

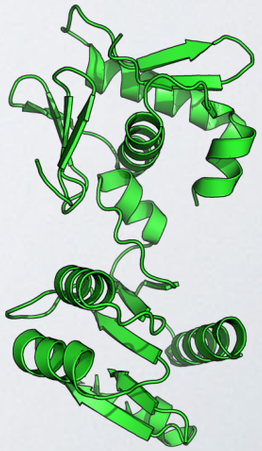
MLHKWMLKNADSQL

GFSKWMLKNADSQL

Template Structures







Sequence-Sequence

SSEARCH (Smith & Waterman)

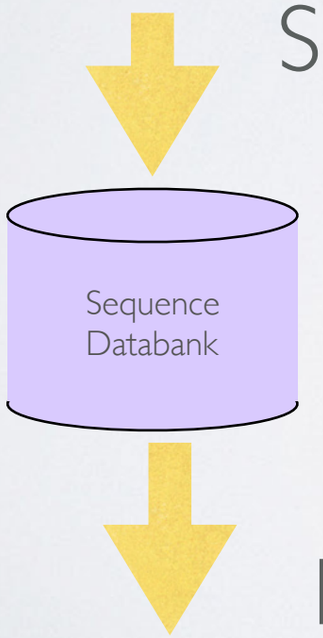
BLAST

HISTORY OF SEQUENCE/ SEQUENCE METHODS

Query Sequence

RHPVVMGNWKLNGSKEMVVDLLNGLNAELEGVTGVDV

Search



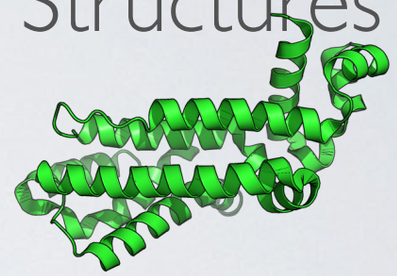
MSA

--RKPIIAGNWKMNGTLAEAVQFVEDVKGHV--PPAD-EVISVVCAPFLFLDRLVQAAD--
AKPQPIAAANWKCNGTTASIEKLVQVFNEH---IS--HDVQCVVAPTFVHIPLVQAKL---
--RHPVVMGNWRLNGSKEMVVDLLNGLNAE--------GVDVAVAPPALFVDLAERTLTEA
--KRPVIANNYRVNGTLAEAVQFVEDVKGHVTVIA--EVISVVCAPFLFLDRLVQAAD---
--KKGIISGNYKIGTLAEAVQFVEDVKGHV--PPAD-EVISVVCAPFLFLDRLVQAAD---

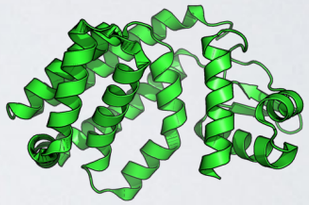
Template Sequences

Template Structures

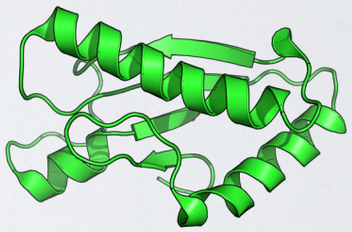
KTVRWMLKNADSQE



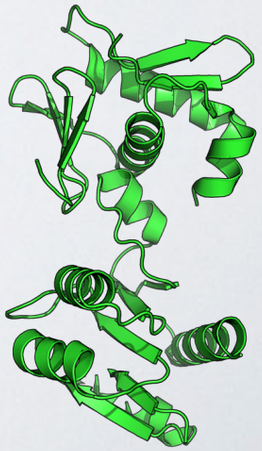
ERHKWMLKNADSQ



LHHKWMLKNADSQL



MLHKWMLKNADSQL



GFSKWMLKNADSQL

HISTORY OF SEQUENCE/ SEQUENCE METHODS

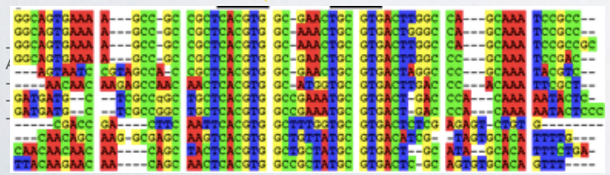
Query Sequence

RHPVVMGNWKLNGSKEMVVDLLNGLNAELEGVTGVDV

Search

Sequence Databank

MSA



Template Sequences

KTVRWMLKNADSQE

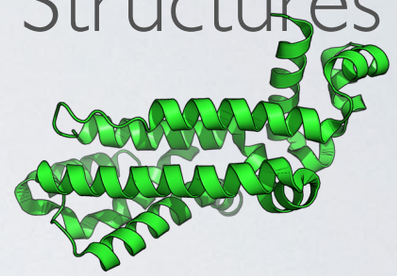
ERHKWMLKNADSQ

LHHKWMLKNADSQL

MLHKWMLKNADSQL

GFSKWMLKNADSQL

Template Structures



HISTORY OF SEQUENCE/ SEQUENCE METHODS

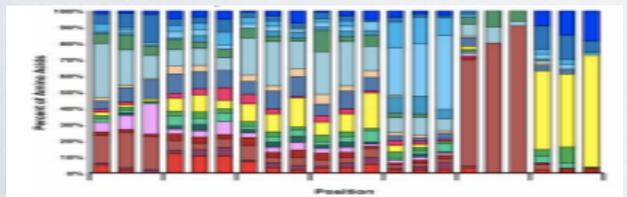
Query Sequence

RHPVVMGNWKLNGSKEMVVDLLNGLNAELEGVTGVDV

Search

Sequence Databank

Profile



Template Sequences

KTVRWMLKNADSQE

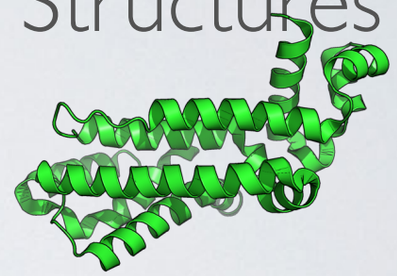
ERHKWMLKNADSQ

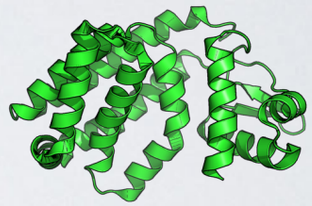
LHHKWMLKNADSQL

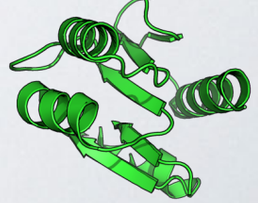
MLHKWMLKNADSQL

GFSKWMLKNADSQL

Template Structures







HISTORY OF SEQUENCE/ SEQUENCE METHODS

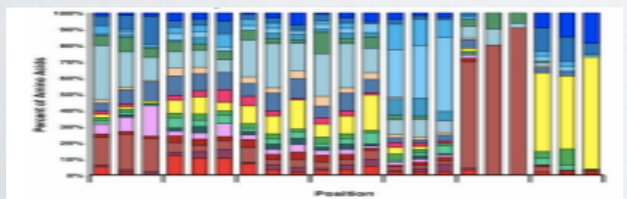
Query Sequence

RHPVVMGNWKLNGSKEMVVDLLNGLNAELEGVTGVDV

Search

Sequence Databank

Profile



Search

Template Sequences

KTVRWMLKNADSQE

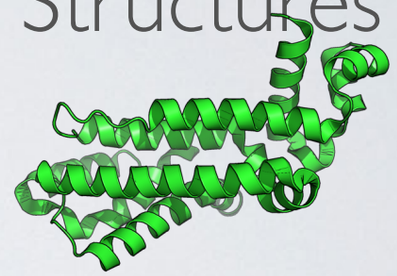
ERHKWMLKNADSQ

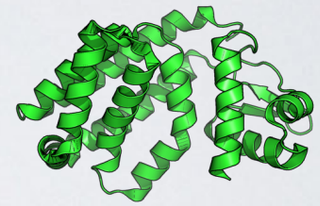
LHHKWMLKNADSQL

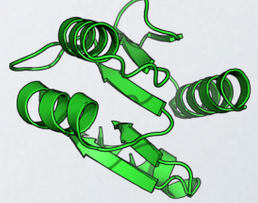
MLHKWMLKNADSQL

GFSKWMLKNADSQL

Template Structures







HISTORY OF SEQUENCE/ SEQUENCE METHODS

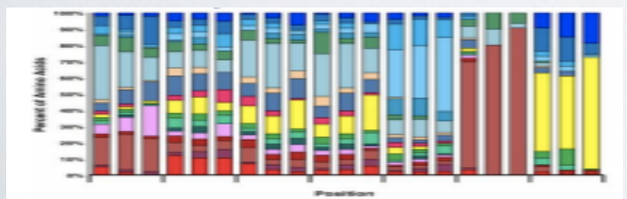
Query Sequence

RHPVVMGNWKLNGSKEMVVDLLNGLNAELEGVTGVDV

Search

Sequence Databank

Profile



Search

Sequence-Profile

PSI-BLAST
SAM
HMMER

Template Sequences

KTVRWMLKNADSQE

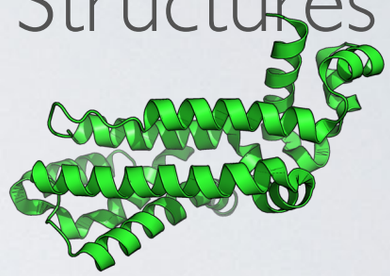
ERHKWMLKNADSQ

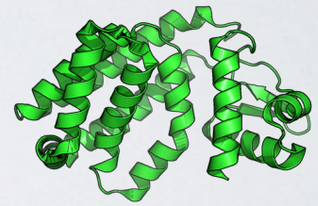
LHHKWMLKNADSQL

MLHKWMLKNADSQL

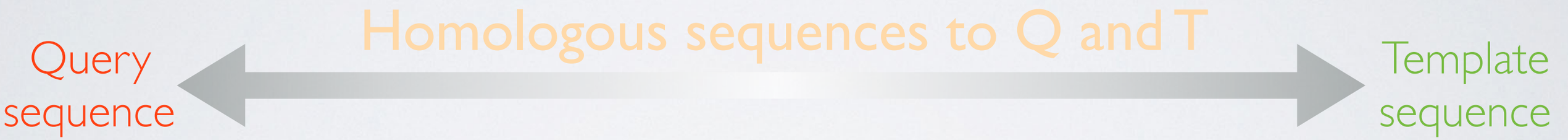
GFSKWMLKNADSQL

Template Structures





TRANSITIVITY



Very distant evolutionary relationship

if query~H and H~target then query~target

HISTORY OF SEQUENCE/ SEQUENCE METHODS

Query Sequence

RHPVVMGNWKLNGSKEMVVDLLNGLNAELEGVTGVDV

Template Sequences

KTVRWMLKNADSQE

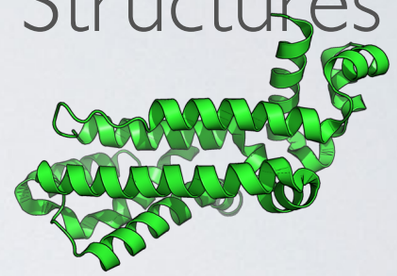
ERHKWMLKNADSQ

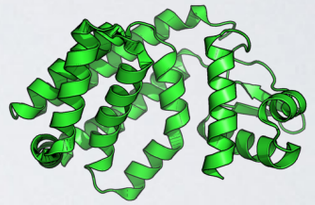
LHHKWMLKNADSQL

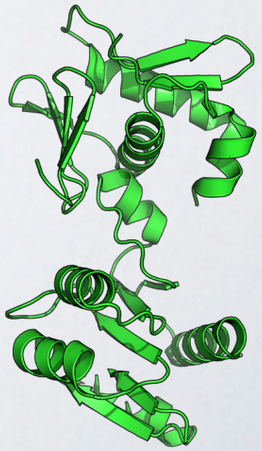
MLHKWMLKNADSQL

GFSKWMLKNADSQL

Template Structures







HISTORY OF SEQUENCE/ SEQUENCE METHODS

Template Sequences

Template Structures

Query Sequence

RHPVVMGNWKLNGSKEMVVDLLNGLNAELEGVTGVDV

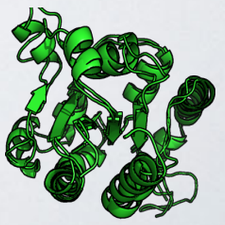
RTVKWMLKNAESQE
KTVKWMLKNTDSQE
KTVRWMLKNADSQE

ERHKWMLKNADSQ
ERHKWMLKNADSQ

LHHKWMLKNADSQL
LHHKWMLKNADSQL

MLHKWMLKNADSQL
-LHKWMLKNA----

GFSKW--KNADSQL
GFSKWMLKNADSQL



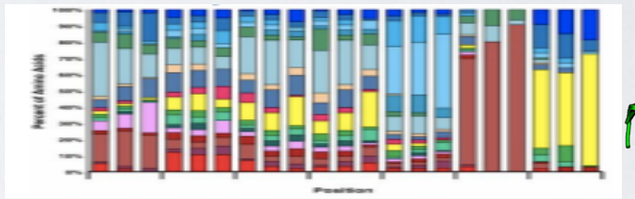
HISTORY OF SEQUENCE/ SEQUENCE METHODS

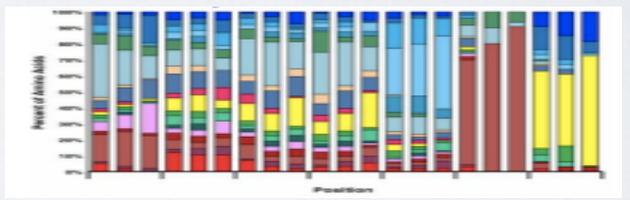
Template Profiles

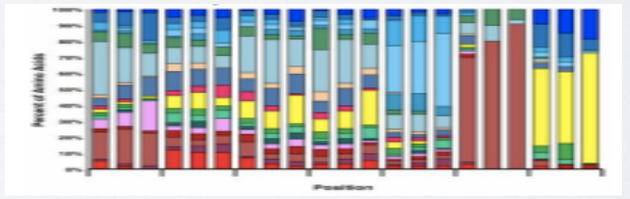
Template Structures

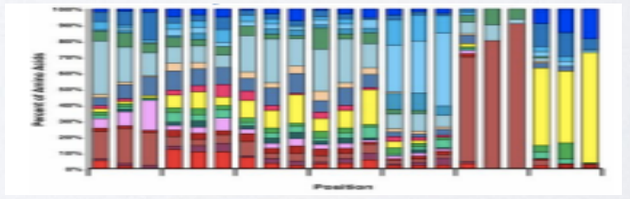
Query Sequence

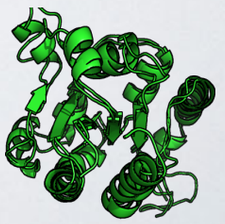
RHPVVMGNWKLNGSKEMVVDLLNGLNAELEGVTGVDV









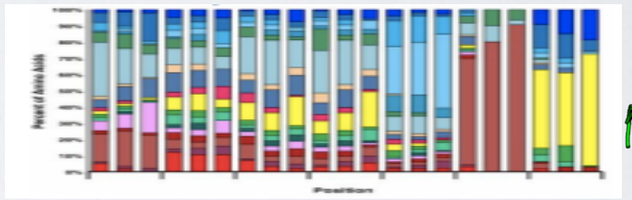


HISTORY OF SEQUENCE/ SEQUENCE METHODS

Template Profiles

Template Structures

Query Sequence

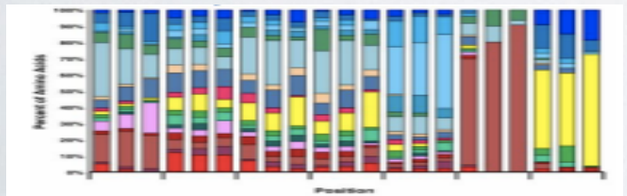


RHPVVMGNWKLNGSKEMVVDLLNGLNAELEGVTGVDV

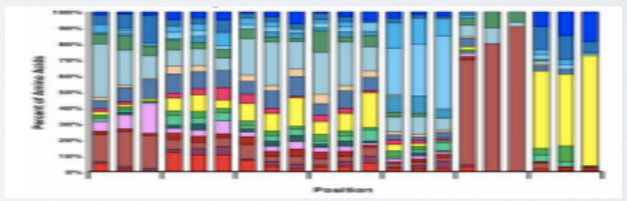
Search

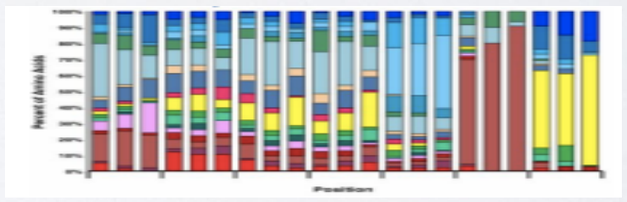
Sequence Databank

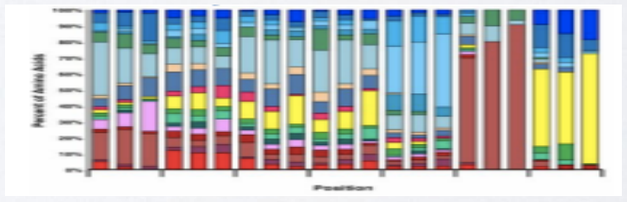
Profile

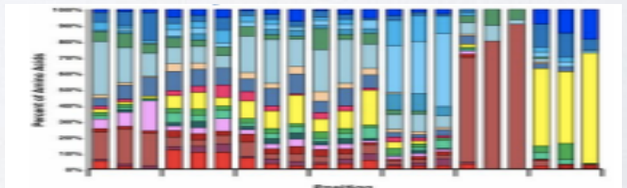


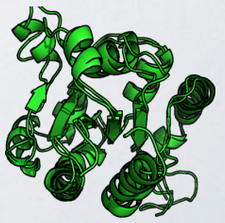
Search











HISTORY OF SEQUENCE/ SEQUENCE METHODS

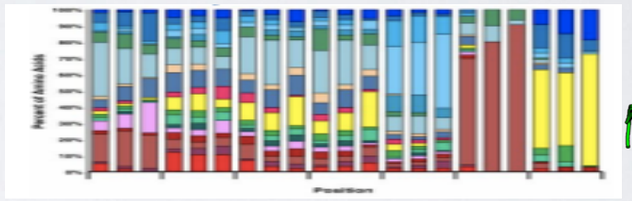
Profile-Profile

Profsim
Compass/Procain
3DPSSM/Phyre/Phyre2
PRC

Template Profiles

Template Structures

Query Sequence



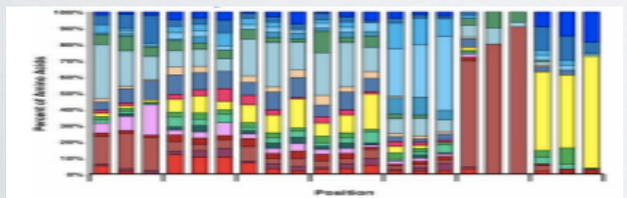
HHsearch

RHPVVMGNWKLNGSKEMVVDLLNGLNAELEGVTGVDV

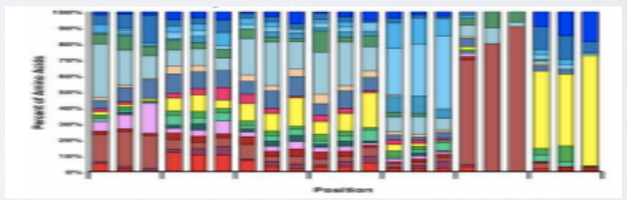
Search

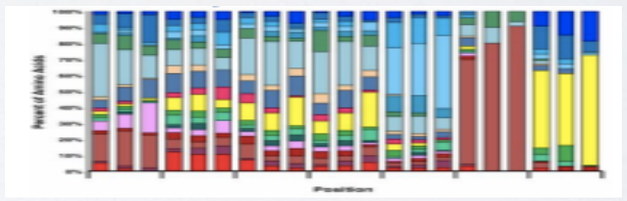
Sequence Databank

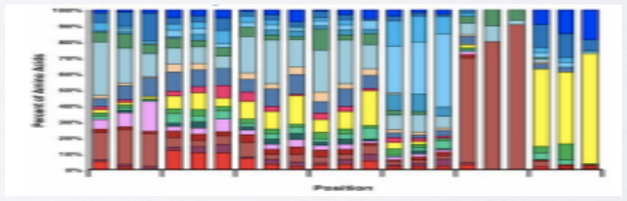
Profile

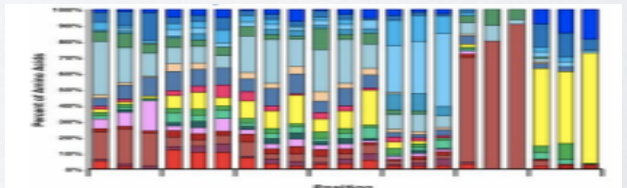


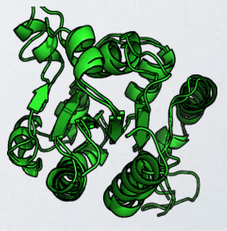
Search











HISTORY OF SEQUENCE/ SEQUENCE METHODS

Query Sequence

RHPVVMGNWKLNGSKEMVVDLLNGLNAELEGVTGVDV

Search

Sequence Databank

Profile

Profile-Profile

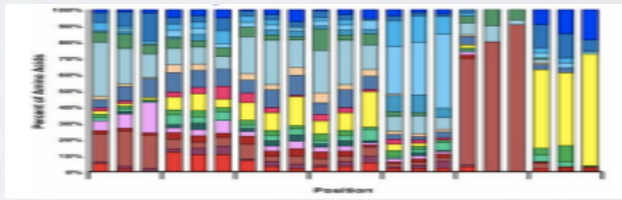
(+ local predictions: S2D and ACC prediction)

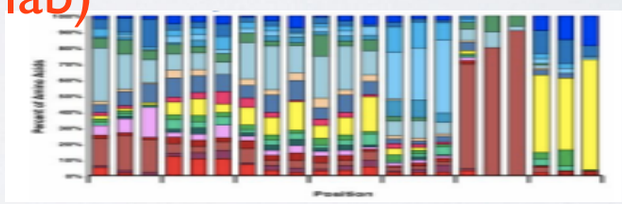
FFAS3D (Godzik lab)
Procain (Grishin lab)
Phyre2 (Sternberg Lab)
RaptorX (Xu lab)
HHsearch [+s2d] (Söding lab)
SPARX (Zhou lab)
Muster (Zhang lab)
Orion (Gelly team)

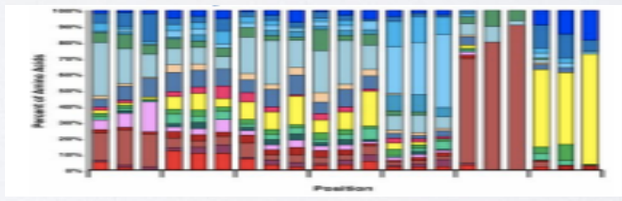
Search

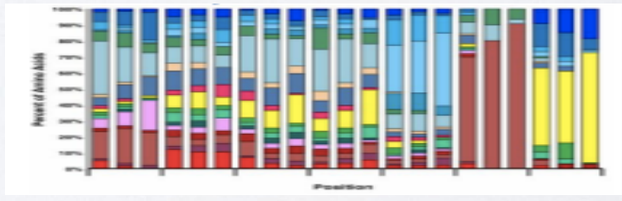
+ local structure predictions

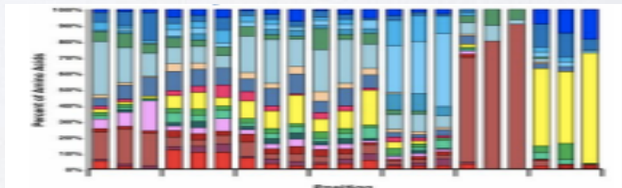
Template Profiles



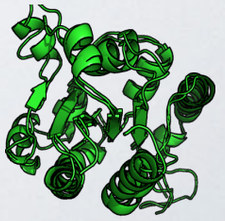






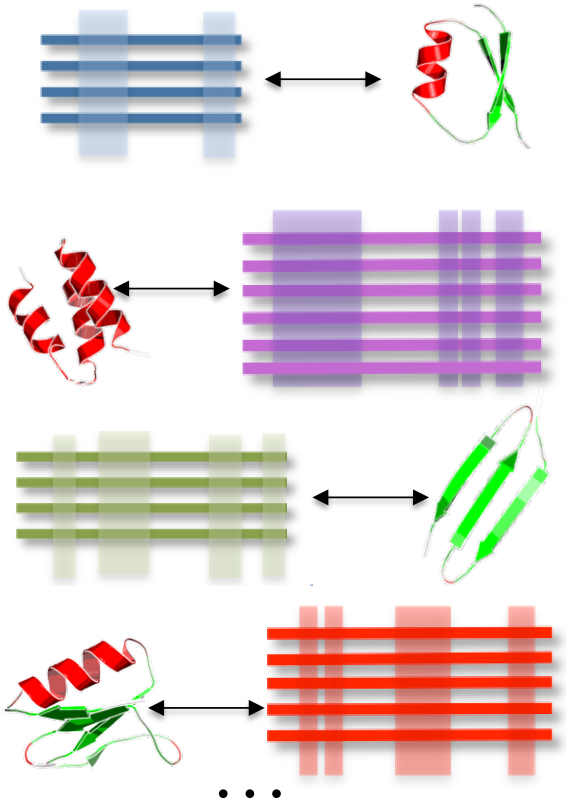


Template Structures



Meet-U 2020

- **Reference database: HOMSTRAD** ***(Mizuguchi K et al, Protein Science 1998)***



We provide a database of about 1000 protein families.

Each family is represented by:

- a master sequence
- a 3D structure (PDB file)
- a multiple sequence alignment
- A SCOP identifier

Meet-U 2020

- **upstream teams**

INPUT
a query sequence

```
ENVIREEAIVRYCATEDIT
```

Meet-U 2020

- **upstream teams**

INPUT
a query sequence

ENVIREEAIVRYCATEDIT

VS

searching homologs

Meet-U 2020

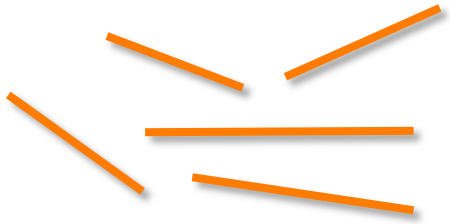
- **upstream teams**

INPUT
a query sequence

ENVIREEAIVRYCATEDIT

VS

searching homologs

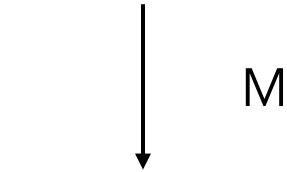


Meet-U 2020

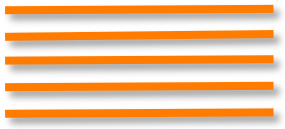
- **upstream teams**

INPUT
a query sequence

`ENVIREEAIVRYCATEDIT`



MSA (MUSCLE, mafft, T-Coffea…)

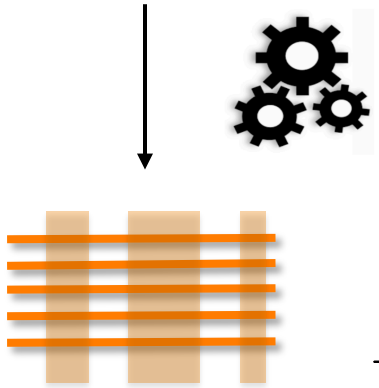


Meet-U 2020

- **upstream teams**

INPUT
a query sequence

ENVIREEAIVRYCATEDIT



building profile

- Full or reduced alphabet
- Added information, like secondary structure, solvent accessibility…
- Sequences can be weighted to reduce redundancy

- **upstream teams**

INPUT
a query sequence

ENVIREEAIVRYCATEDIT

Meet-U 2020

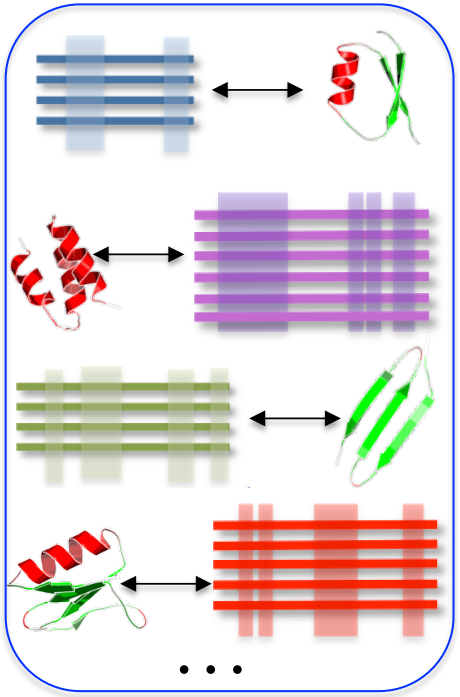
- **upstream teams**

INPUT
a query sequence

ENVIREEAIVRYCATEDIT

building profile

sequences-structures DB



Meet-U 2020

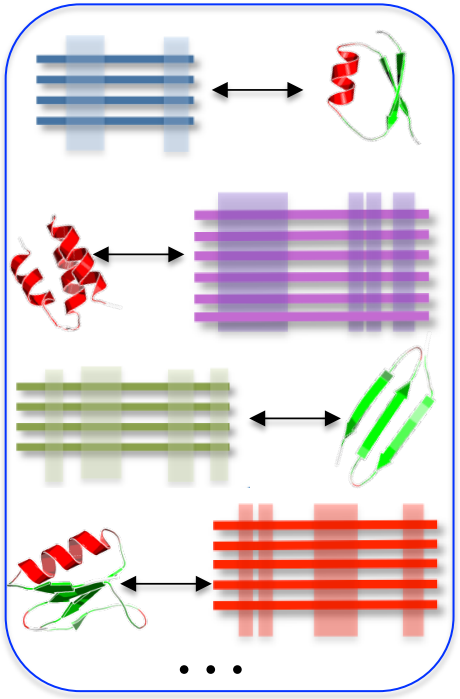
- **upstream teams**

INPUT
a query sequence

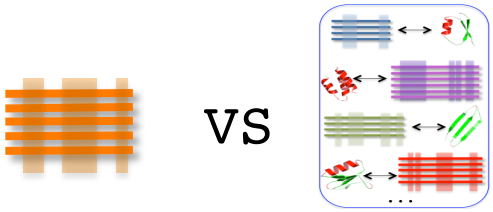
ENVIREEAIVRYCATEDIT

building profile

sequences-structures DB



profile-profile comparison



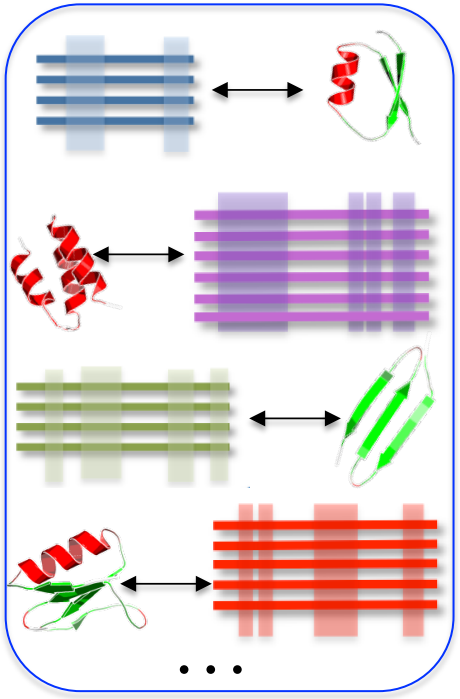
Meet-U 2020

- **upstream teams**

INPUT
a query sequence

ENVIREEAIVRYCATEDIT

sequences-structures DB



building profile

profile-profile comparison

vs

x score #1

x score #2

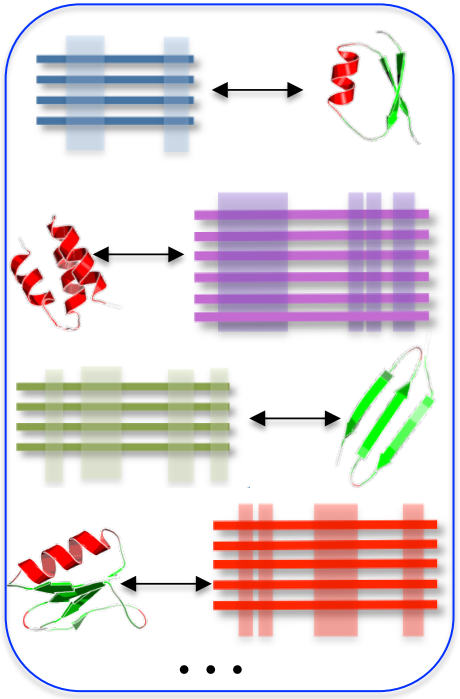
Meet-U 2020

- **upstream teams**

INPUT
a query sequence

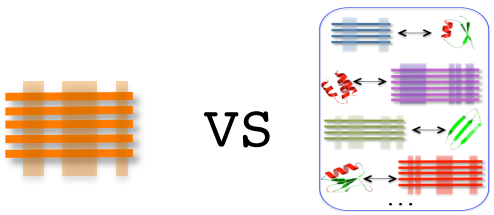
ENVIREEAIVRYCATEDIT

sequences-structures DB



building profile

profile-profile comparison



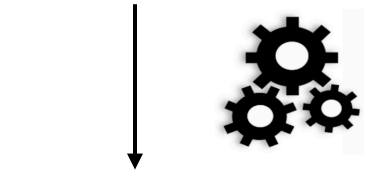
- Alignment algorithm (global, local, hybrid)
- Scoring function (dot product…)

Meet-U 2020

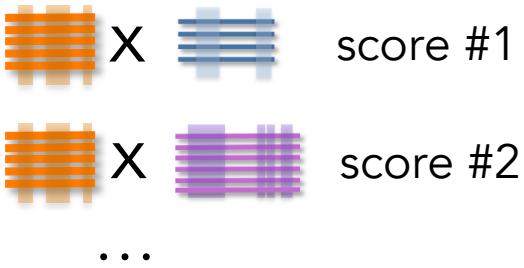
- **upstream teams**

INPUT
a query sequence

ENVIREEAIVRYCATEDIT



OUTPUT
N profil-profile alignments and their corresponding scores



Meet-U 2020

- **upstream teams**

INPUT
a query sequence

ENVIREEAIVRYCATEDIT

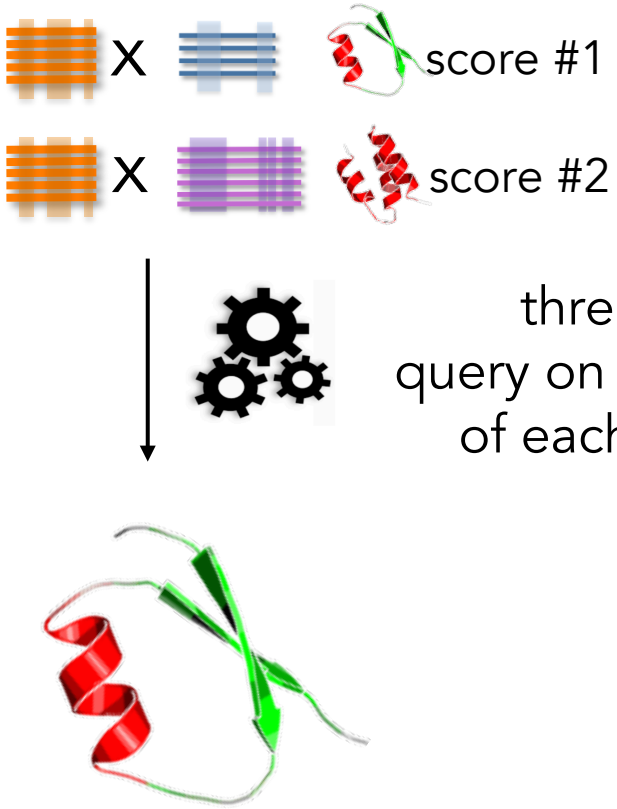
OUTPUT
N profil-profile alignments and their corresponding scores

- **downstream teams**

Meet-U 2020

- **downstream teams**

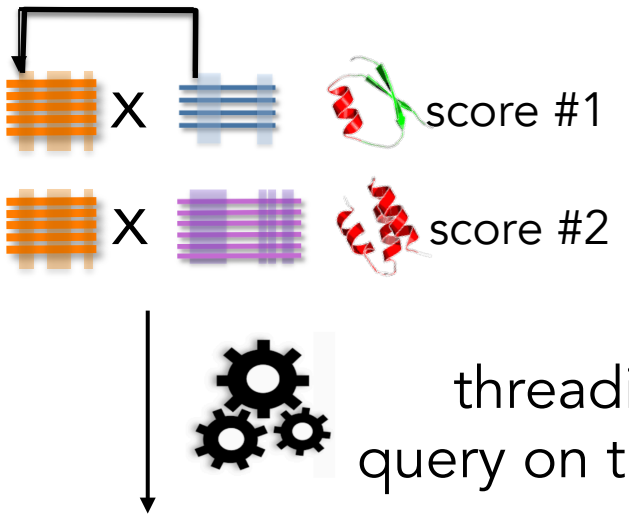
INPUT
N profile-profile alignments and their corresponding scores



Meet-U 2020

- **downstream teams**

INPUT
N profile-profile alignments and their corresponding scores

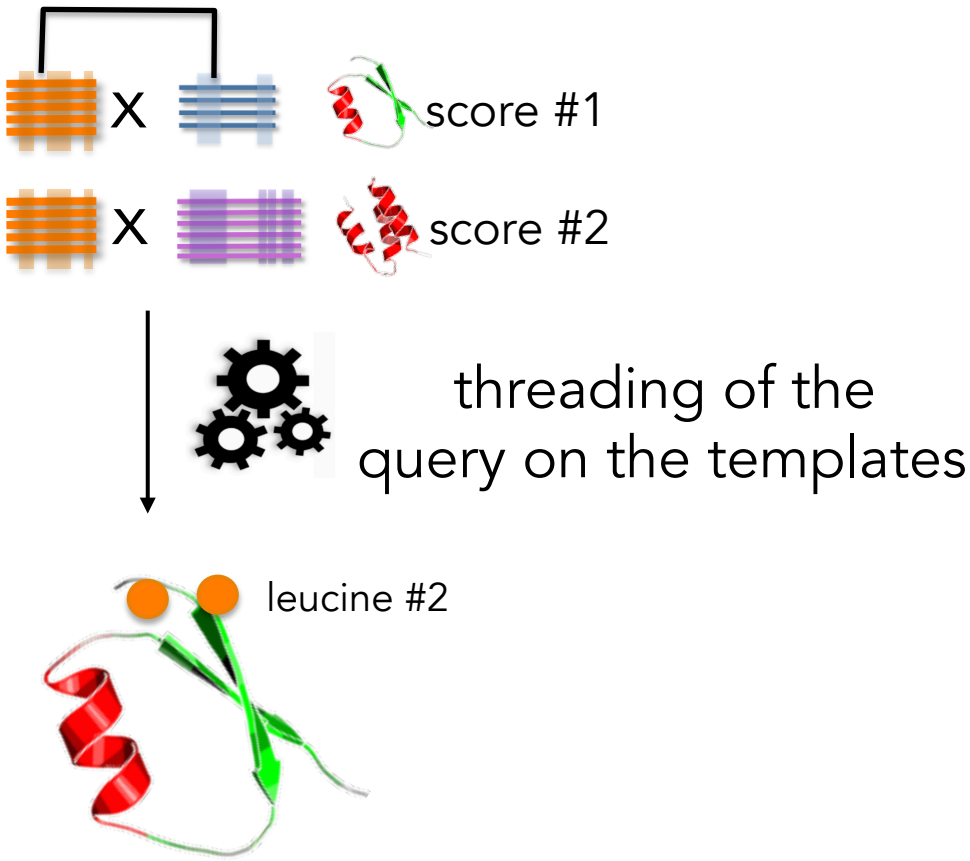


threading of the query on the templates

Meet-U 2020

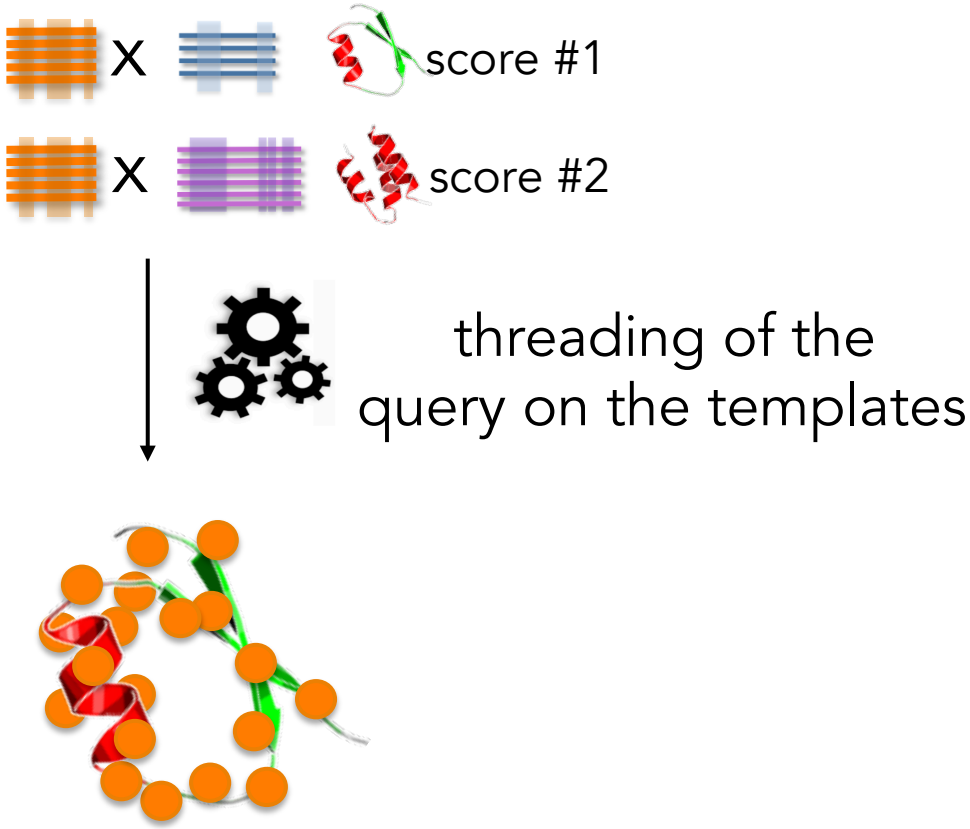
- **downstream teams**

INPUT
N profile-profile alignments and their corresponding scores



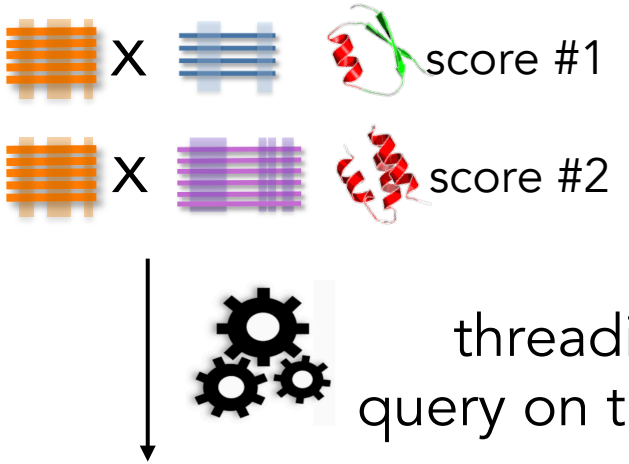
- **downstream teams**

INPUT
N profile-profile alignments and their corresponding scores

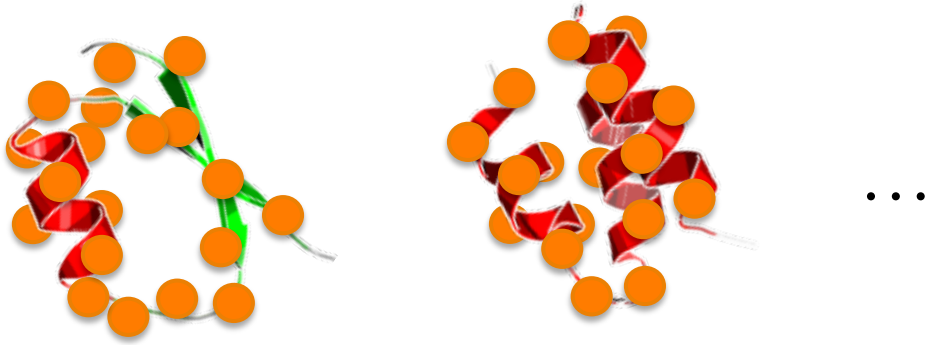


- **downstream teams**

INPUT
N profile-profile alignments and their corresponding scores



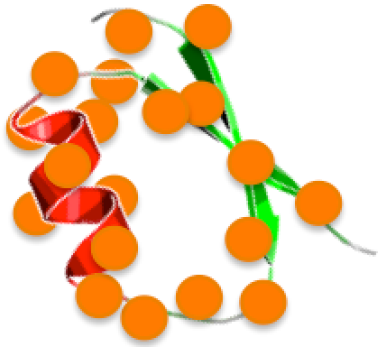
N models



- **downstream teams**

...

assessment of the quality of the models



model #1

SCORING FUNCTION

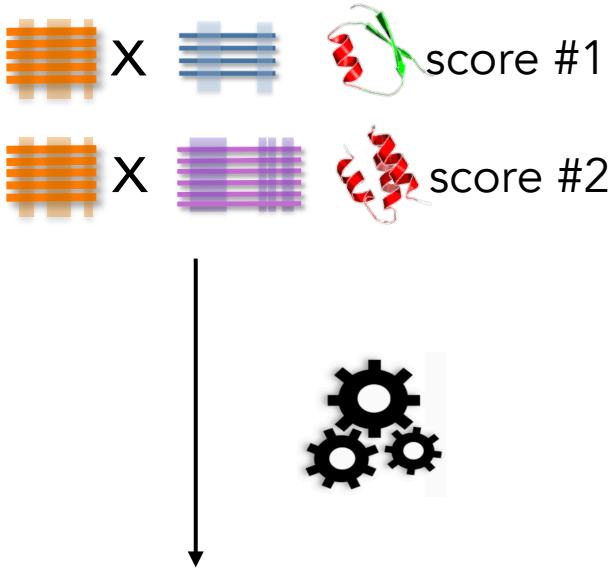
Pairwise distance-dependent DOPE score

Plus information coming from the alignment, coevolution signals, secondary structure & solvent accessibility consistency, physical terms…etc

Meet-U 2020

- **downstream teams**

INPUT
N profile-profile alignments and their corresponding scores



OUTPUT

X models and their corresponding scores

model #46
TOP1

model #38
TOP2

...

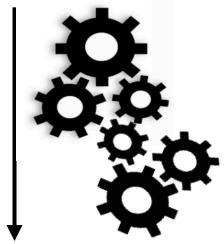
model #72
TOPX

Meet-U 2020

INPUT
query sequence

ENVIREEAIVRYCATEDIT

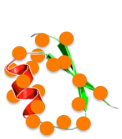
FASTA

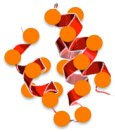


- **Upstream part**
- **Downstream part**

OUTPUT

X models and their corresponding scores





model #46
TOP1

model #38
TOP2

...

model #72
TOPX

- PDB files containing model coordinates
- One 2-column file with models scores

Meet-U 2020

- Benchmark to train(?), test/evaluate your method
  - 1000 examples with different levels of difficulties

seqtest #1 MTKRVELALIDKFPMFLAVITRYCESTFINI

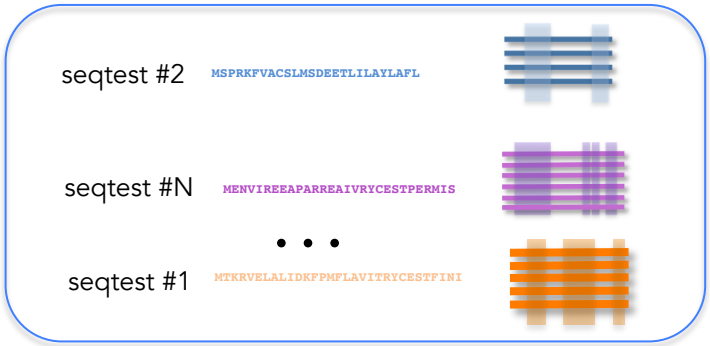
seqtest #2 MSPRKFVACSLMSDEETLILAYLAFL

. . .

seqtest #N MENVIREEAPARISSAMEDIENIVREAIVRYCESTPERMIS

sequences homologous to the different families of the database

+



File with the solutions (i.e. the true associations sequences-families)

WHEN & WHERE ?

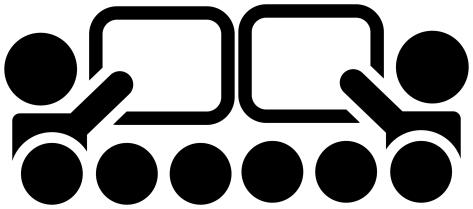
09/27
Opening Session
@SU

« practical »
sessions of
3h

11/08
Half-way forum
@SU

« practical »
sessions of
3h

01/08?
Meeting Day
@Orsay



Flipped symposium

It's time to score!

Then, it is our turn!

Evaluations - criteria

5 criteria evaluated by the pedagogical team

- program (quality, property of the code, lisibility, usability)
- strategy: originality, relevance, risk
- result analysis: relevance, critical mind
- organization, participation, team work
- report (report 10 pages + digest 2 pages), presentations

3 criteria evaluated by the jury members

- presentation
- strategy: originality, relevance, risk
- result analysis: relevance, critical mind

Tâches principales

- **Formation des groupes (4, 5 max) en fonction des compétences de chacun**
  - ➢ **qui fait quoi ? définition des rôles**
- Recherche bibliographique et appropriation des concepts
  - ➢ rapport synthétique, présentation à l'équipe
- Définition de la stratégie et des choix méthodologiques
  - ➢ rapport détaillé, envisager les risques
- Implémentation de l'outil
  - ➢ code et documentation (README & commentaires)
- Validation de l'outil
  - ➢ applications à des cas tests, justification des résultats
- **Présentation de l'outil et des résultats, dissémination**
  - ➢ **présentation orale devant le jury**

Déroulé du projet

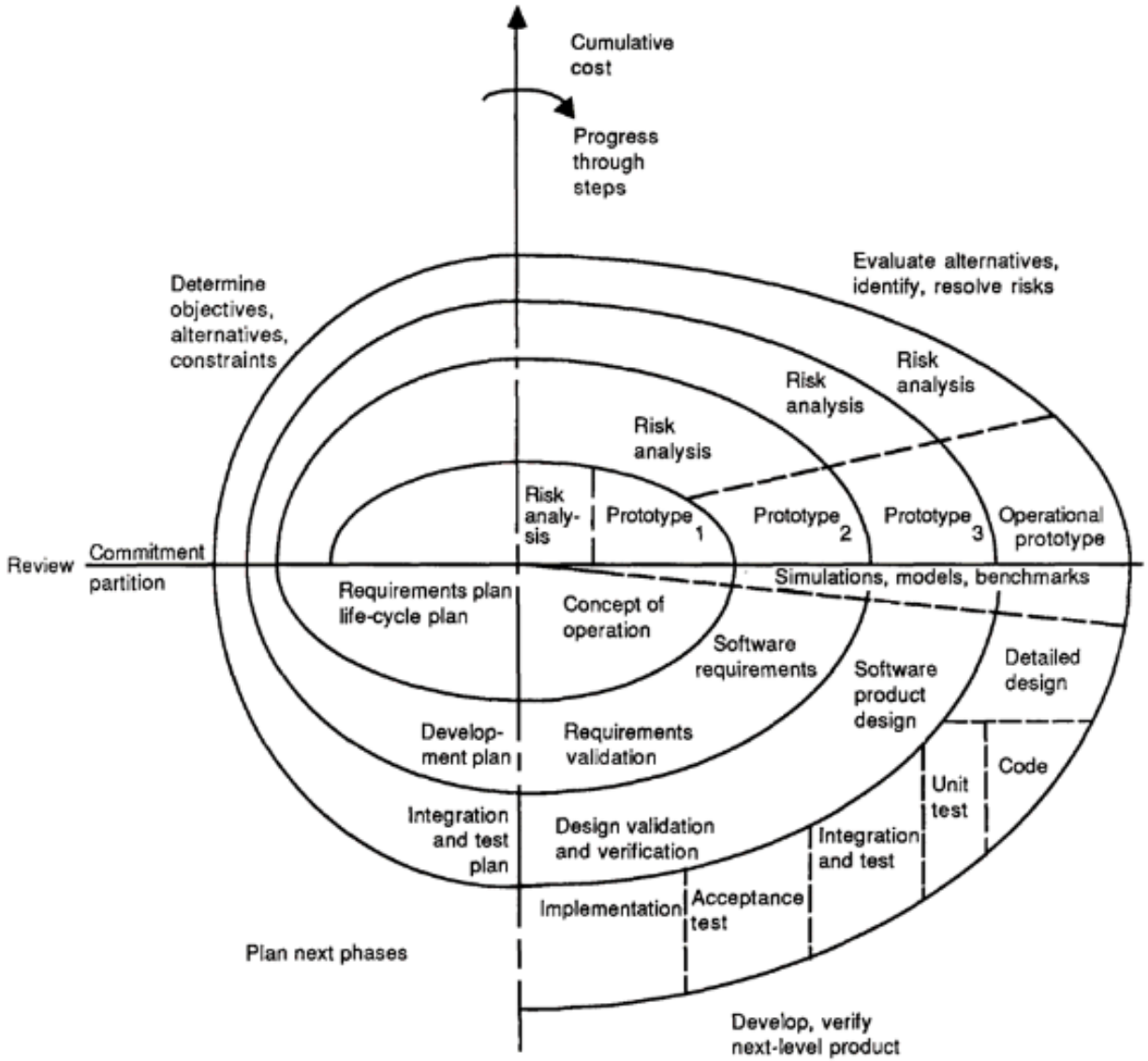


Figure 2. Spiral model of the software process.

Déroulé du projet

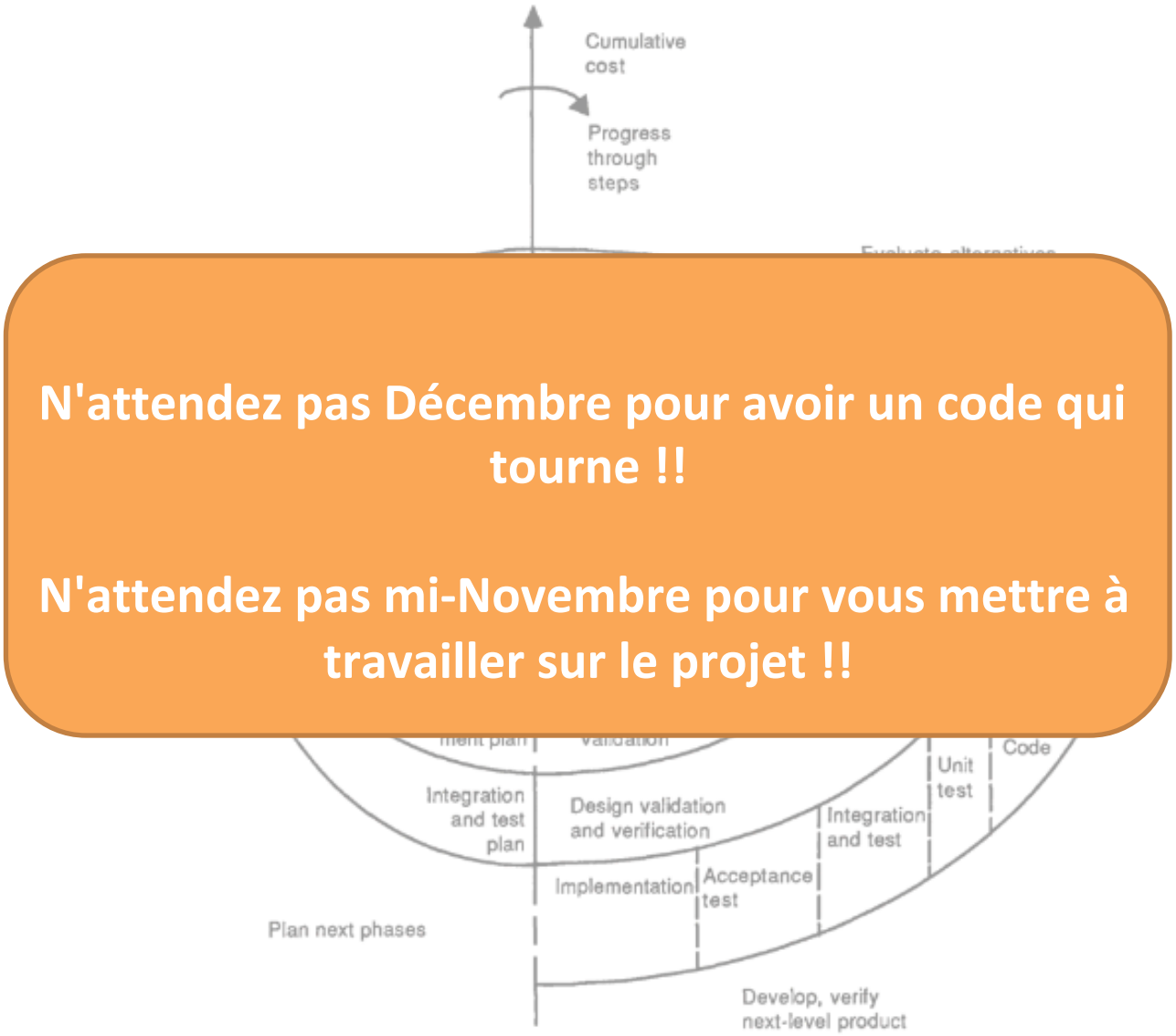


Figure 2. Spiral model of the software process.

N'attendez pas Décembre pour avoir un code qui tourne !!

N'attendez pas mi-Novembre pour vous mettre à travailler sur le projet !!

Rôles dans le projet

- ✓ **Manager** : gestion des ressources et du temps, animation meetings
- ✓ **Expert(s) technique(s)** : programmation (implémentation et validation)
- ✓ **Expert(s) veille scientifique** : recherche bibliographique & propositions
- ✓ **Expert(s) déliverables** : rédaction des rapports, présentation

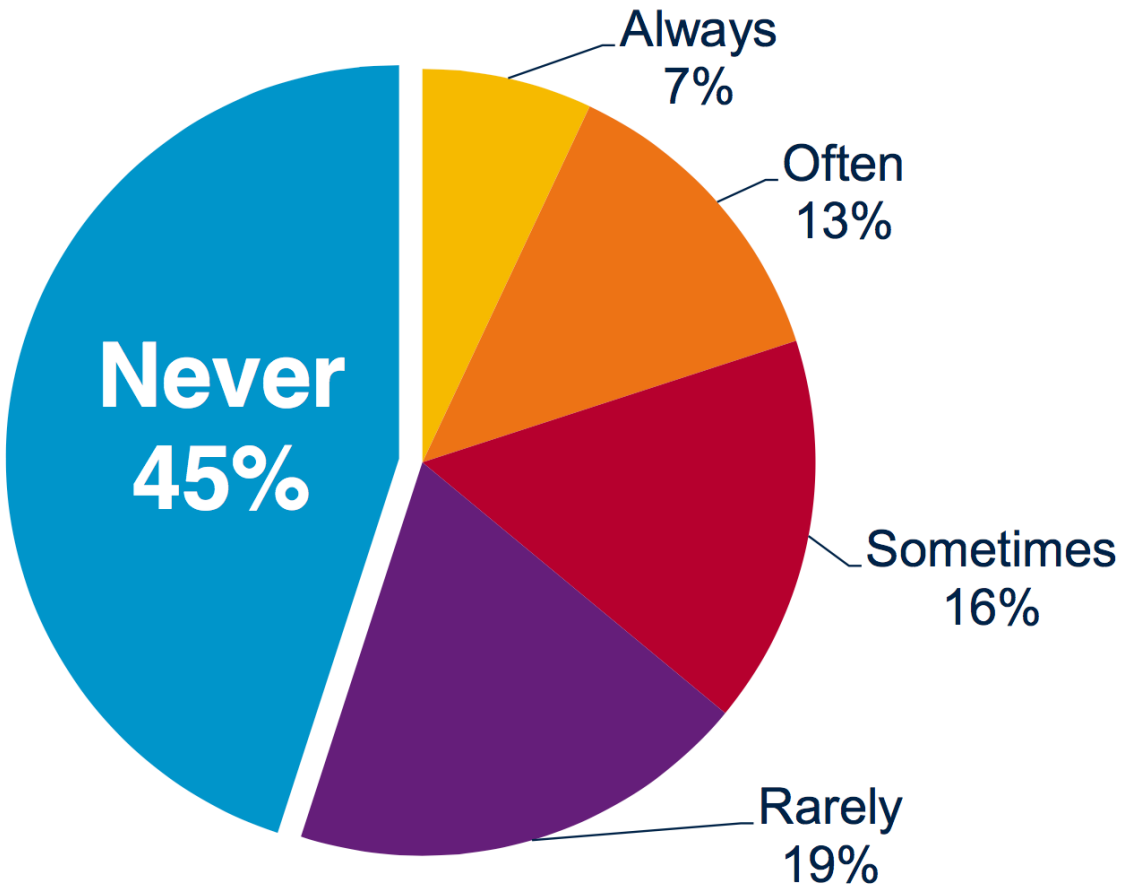
Tous ces rôles s'exerceront tout au long du projet.

Les experts doivent échanger !

Chacun peut jouer plusieurs rôles, mais il est important de bien définir les responsabilités associées.

Bonnes pratiques

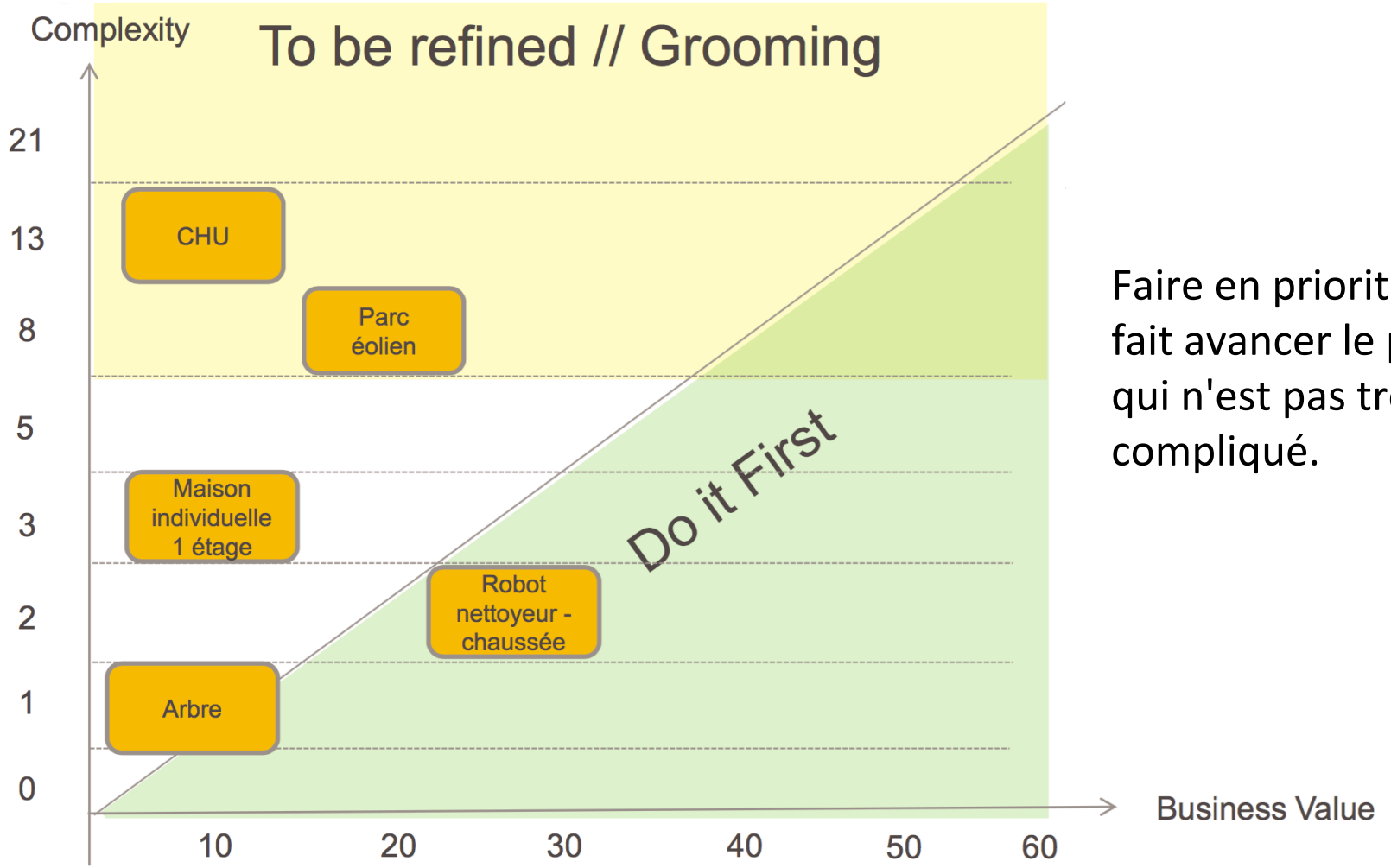
Près de la moitié des fonctions développées ne sont jamais utilisées.



Bonnes pratiques

- ✓ Il est impossible de rassembler/connaître tout ce qui sera requis/utile dès le début du projet.
  - ➢ **N'essayez pas de concevoir le projet parfaitement et complètement avant de commencer à implémenter**
- ✓ Quels que soient votre stratégie et vos choix initiaux, ils seront amenés à changer au cours du projet.
  - ➢ **Soyez flexibles ! Autorisez-vous des allers-retours, implémenter vos solutions de manière modulaire, code facile à enrichir/adapter**
- ✓ Il y aura toujours plus de choses à faire que le temps et les ressources ne le permettent.
  - ➢ **Ne vous fixez pas d'objectifs inatteignables, soyez réalistes et faites des compromis.**

Bonnes pratiques



Faire en priorité ce qui fait avancer le projet et qui n'est pas trop compliqué.

Bonnes pratiques

	A FAIRE	EN COURS	FAIT
CRITIQUE		Pb1	
MAJEUR			Pb5 Pb4
MINEUR	Pb6	Pb7	

	PAS URGENT	URGENT
CRITIQUE		
MAJEUR		
MINEUR		

To do for next time

- creation of your GitHub account and watch the tutorial
- choice of the strategy – (and split into tasks)
- definition of the roles of each team member
- start implementing your part
- preparation of 2-3 slides (informal) to present your strategy

Conseil du jour

LA PERFECTION EST L'ENNEMI DU BIEN…