L'informatique à l'INRA

Journée Enseignement et recherche en informatique "hors les murs"

SPECIF Campus
Télécom Paris Tech, Paris – 24 novembre 2016

INRA

Institut national de la recherche agronomique

1er institut de recherche agronomique en Europe,
2è en sciences agricoles dans le monde

Des recherches finalisées pour

- une alimentation saine et de qualité,
- une agriculture durable,
- un environnement préservé et valorisé.

Production académique en 2015

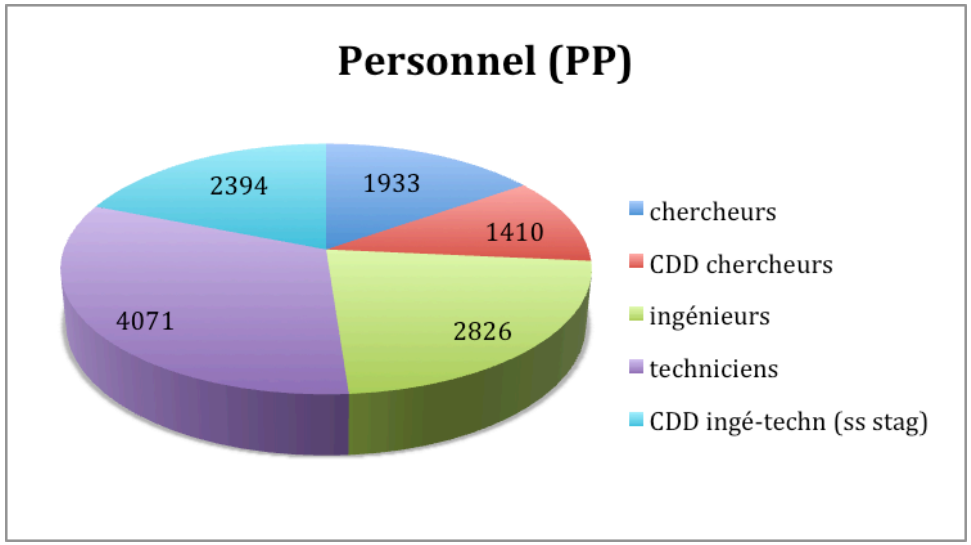
- 4081 publications (Source WoS)
- 55 brevets (354 au total)
- 14 nouvelles variétés végétales

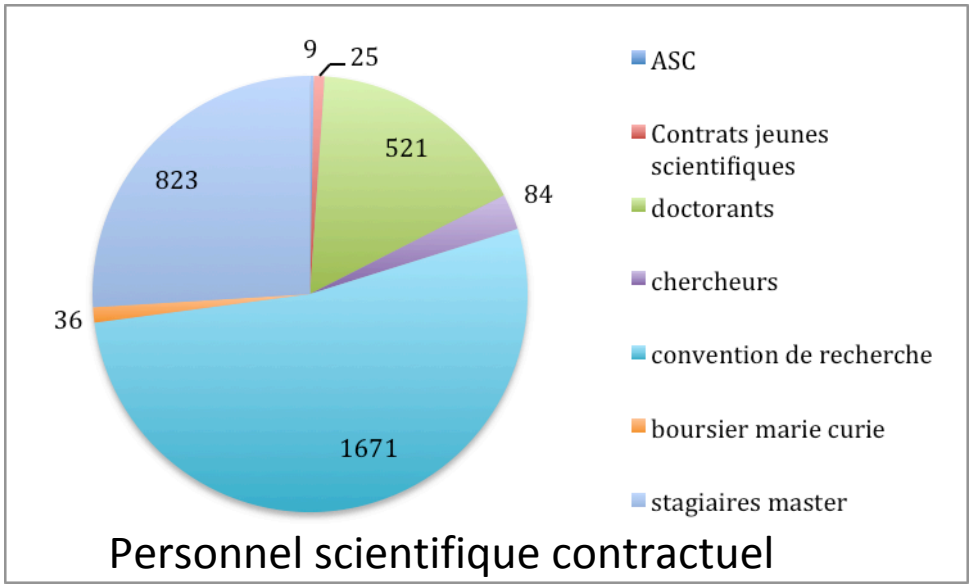
Quelques chiffres

250 unités de recherche et **48 unités expérimentales** dans **13 départements scientifiques**
150 unités mixtes (écoles, universités, CNRS, INRIA, CIRAD, IRD, ...)

17 centres régionaux et 1 siège (Paris)
150 sites géographiques

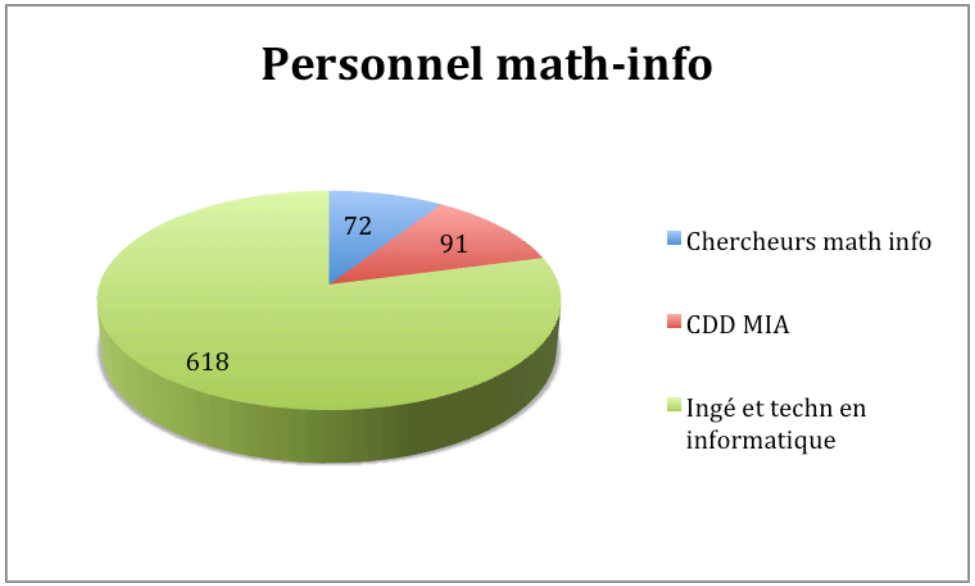
10 240 titulaires en personnes physiques hors stagiaires (bilan social 2014)
3169 scientifiques contractuels (5330 au total)

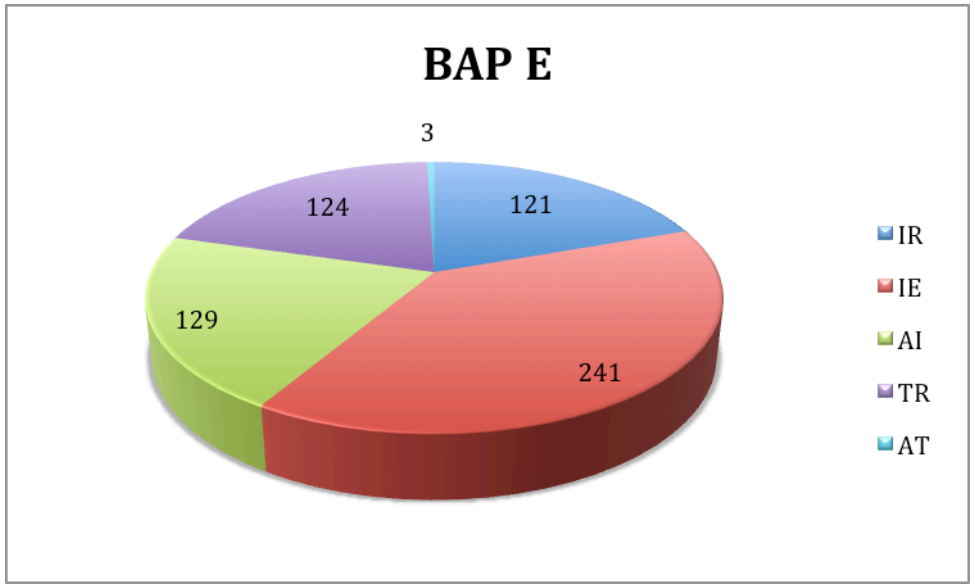




Personnel math-informatique

4% du total des chercheurs titulaires (72 dans la CSS MBIA)
11% du total des ingénieurs et techniciens titulaires (618 BAP E)





13 départements scientifiques

8 metaprogrammes transversaux

Animal	Génétique animale Santé animale Physiologie animale et systèmes d'élevage
Microbiologie et alimentation	Microbiologie et chaîne alimentaire Alimentation humaine
Plante	Biologie et amélioration des plantes Santé des plantes et environnement
Transformation	Caractérisation et élaboration des produits issus de l'agriculture
Environnement	Écologie des forêts, prairies et milieux aquatiques Environnement et agronomie
SHS	Sciences pour l'action et le développement Sciences sociales, agriculture et alimentation, espace et environnement
Math-info	Mathématiques et informatique appliquées

Mathématiques et informatique

Un continuum de recherche, des mathématiques appliquées à l'informatique, pour les Sciences du Vivant.

Un portail commun. Une trentaine de laboratoires décrits.

Le département MIA

Mathématique et Informatique appliquées

- *M*ener des **recherches en maths-info et à leur interface** avec d'autres disciplines, pour répondre aux grands enjeux de la recherche en sciences du vivant et de l'environnement;
- *A***ccompagner la transition numérique et le développement des maths-infos** à l'INRA

127 titulaires, 91 contractuels dans 6 unités de recherche

- **MaIAGE** (UR *Mathématiques, Informatique, du Génome à l'Environnement*) à Jouy-en-Josas (UPSay),
- **MIA-Paris** (UMR *Mathématiques et Informatique appliquées*) UPSay,
- **LaMME** (USC *Laboratoire de Mathématiques et Modélisation d'Évry*) UPSay,
- **MIAT** (UR *Mathématiques et Informatique appliquées*, Toulouse)
- **BioSP** (UR *Biostatistique et Processus Spatiaux*) à Avignon
- **Mistea** (UMR *Mathématiques, Informatique et STatistique pour l'Environnement et l'Agronomie*) à Montpellier

+ **IFB-core**, UMR de service, noeud national de l'IFB (*Institut Français de Bioinformatique*)

+ **InGeNum**, UR appui

Méthodes

Axes méthodologiques, déclinaison en informatique

1. De la **donnée** à la connaissance
Extraction, représentation des connaissances et raisonnement. Intelligence artificielle, traitement automatique de la langue, ontologies, analyse sémantique, apprentissage

2. Modélisation intégrative de systèmes complexes et multi-échelles
Couplage de modèles multi-échelles, variables discrètes et continues, systèmes dynamiques

3. Optimisation et **conception** à partir de modèles
Informatique : Simulation, Optimisation

Domaines d'intérêt

1 : Biologie computationnelle, systémique et synthétique

- génomique et métagénomique
- réseaux de régulation
- lien génotype-phénotype
- sélection, biologie de synthèse

2 : Biologie des populations, écologie, épidémiologie

- modélisation dynamique d'écosystèmes microbiens
- processus de propagation en écologie, santé des plantes et des animaux
- épidémiosurveillance, contrôle de bioagresseurs, biodiversité

3 : Agriculture et environnement numériques

- interactions génotype-environnement
- données climatiques à fines échelles temporelles et spatiales
- risques agro-environnementaux pour l'aide à la décision publique

Animation de la recherche et du support à la recherche en *math-info*

- **15 réseaux scientifiques** nationaux soutenus par le département MIA
- **18 CATI scientifiques nationaux** : groupes d'ingénieurs et techniciens en Calcul Scientifique et Informatique pour faciliter et mutualiser le soutien aux équipes de recherche.
- **7 Pepi**, Partage d'Expériences et de Pratiques en Informatique : réseaux métiers
- **Délégation à la transition numérique, unité IngéNum** : coordination de la transition numérique en recherche
- **Cellule bioinformatique** : coordination de la bioinformatique
- **7 plateformes** : e-infrastructures scientifiques partagées pour le traitement et l'analyse de données, la modélisation et la simulation

15 réseaux scientifiques nationaux math-info

IN OVIVE : Intégration de sources/masses de données hétérogènes et ontologie en sciences du vivant, de l'agronomie et de l'agro-alimentaire. Intelligence artificielle

AIGM : Modélisation spatio-temporelle sur graphes et approximations informatique pour l'inférence dans les modèles graphiques, déterministes ou stochastiques

NETBIO : Inférence de réseaux biologiques, modèles graphiques probabilistes et des méthodes d'inférence statistiques

Optim : Optimisation pour les Sciences de la Vie, discrète ou continue, linéaire ou non, convexe, combinatoire, exacte ou approchée

PAYOTE : Modélisation des paysages et territoires agricoles pour la simulation et l'analyse des processus environnementaux

Une recherche collaborative au niveau national et international

Thomas Schiex reçoit le EurAI Fellows, pour ses travau exceptionnels au service de l'Intelligence Artificielle.

CATI, Centre automatisé de traitement de l'information

Des CATI "plateforme" et des CATI "réseau"
Des ingénieurs et techniciens, mais aussi des scientifiques (support dans les équipes de recherche)

Cati scientifiques en 5 thèmes :

- bioinformatique (7),
- modélisation (2),
- expérience et observation (4),
- intelligence artificielle (1),
- calcul scientifique (1)

Plateformes et e-infrastructures partagés

90 actifs stratégiques en informatique (soutien à la production scientifique)

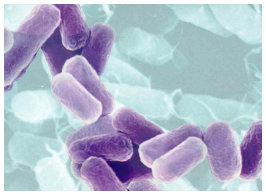
- Plateformes –omiques, de systématique et de ressources génétiques
- Expérimentation et d'observation
- Modélisation
- Imagerie
- Analyse textuelle
- Calcul scientifique

Plateformes du périmètre MIA

- **Plateformes bioinformatiques** : plateformes MIA : Migale, GenoToul (+ URGI, SIGENAE, South-Green)
- Modélisation et simulation des **agro-écosystèmes** : plateforme MIA RECORD
- **Extraction d'information** et analyse textuelle : plateforme MIA Alvis, (+ CoreText)

Démarches de collaboration à l'interface

Deux exemples illustratifs



Ecosystèmes microbiens et alimentation

Un exemple de coconception de service bioinformatique en ligne alliant *recherches* et *d'ingénierie* en biologie, en informatique et en bioinformatique.

Reconstruction des réseaux de régulation impliqués dans le développement de la plante

Un exemple de dialogue *interdisciplinaire en informatique et en biologie*, médiatisé par un éditeur d'annotation.

OntoBiotope, une application TDM pour la microbiologie

Mieux connaître les microorganismes et l'adaptation à leur milieu

Un enjeu pour la recherche et l'industrie : santé, agro-alimentaire, environnement.

Des développements majeurs des technologies moléculaires et des sciences de l'information

- **Déluge de données expérimentales, modélisation des mécanismes.**
- **Déluge de documents en texte libre, articles, brevets, bases de données.**

Des infrastructures bioinformatiques

pour l'analyse intégrée de l'information, multi-sources, multi-espèces, multi-échelle.

OntoBiotope

Extraire des textes toute l'information sur les habitats des microorganismes et la **normaliser** pour l'analyser avec l'information expérimentale.

Un exemple d'application d'OntoBiotope : *Food Microbiome*

Projet INRA - CNIEL (*Centre National Interprofessionnel de l'Economie Laitière*)

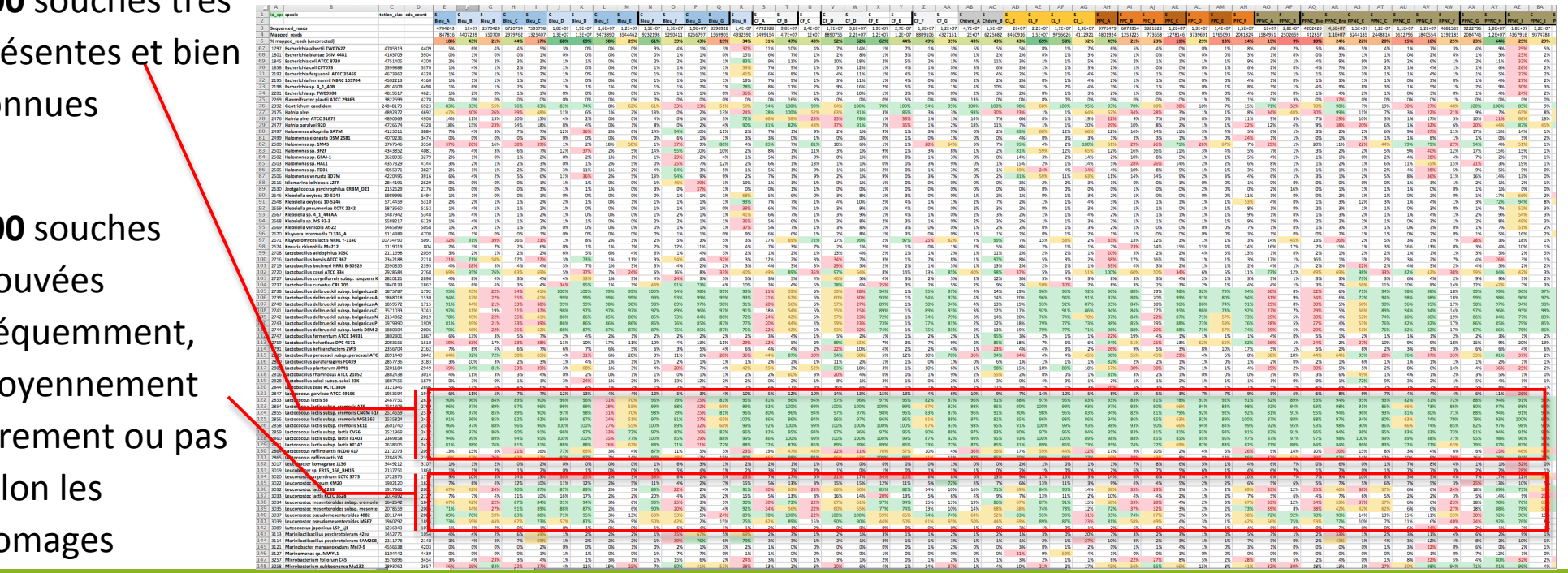
Identifier les bactéries dans les produits laitiers, comprendre et contrôler leur présence, améliorer la qualité des produits, innover.

Analyse *métagénomique* de centaines d'échantillon de fromages français et italiens

300 souches très présentes et bien connues

100 souches trouvées fréquemment, moyennement rarement ou pas selon les fromages

Sur 400 souches



Des souches mystérieuses

Pour chaque souche identifiée

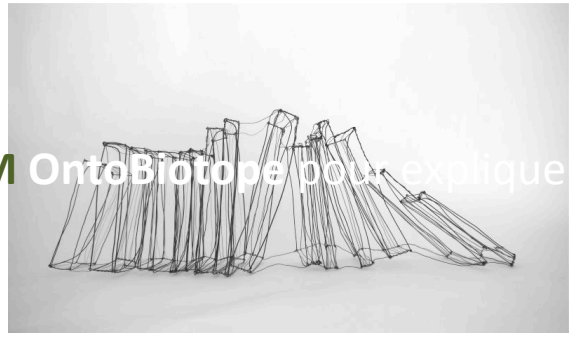
Est-ce vraisemblable ? A-t-elle déjà été trouvée dans ce fromage ? Une bactérie de la même famille ? Dans un milieu similaire ? Des explications connues sur leur présence dans le fromage ?

Psychrobacter aquimaris
ER15_174_BHI7

Gorgonzola, roquefort, époisse, toscanello, st nectaire (très présente), tomme, bleu

Présente dans tous les échantillons ou certains, ou complètement absente. Selon les fromages

L'application **TDM OntoBiotope** pour expliquer leur présence



© Myriam Louvel Paoli 2015

bibliome.jouy.inra.fr

INRA SCIENCE & IMPACT

OntoBiotope Database

Welcome to the Ontobiotope database. You can browse through bacteria and their habitats found in over 700,000 PubMed abstracts.

You can start exploring the data either by bacteria taxonomy, or by the bacteria habitat ontology.

Alvis Search Engine

Psychrobacter aquimaris

Search habitat

Bacteria | Proteobacteria | Gammaproteobacteria | Pseudomonadales | Moraxellaceae | Psychrobacter | Psychrobacter aquimaris

bacteria habitat

Psychrobacter aquimaris X bacteria habitat

Drag a column and drop it here to group by that column

Title	Taxon	Habitat
Psychrobacter aquimaris sp. nov. and Psychrobacter	Psychrobacter aquimaris	marine environment [sea water of the South Sea]

Psychrobacter aquimaris

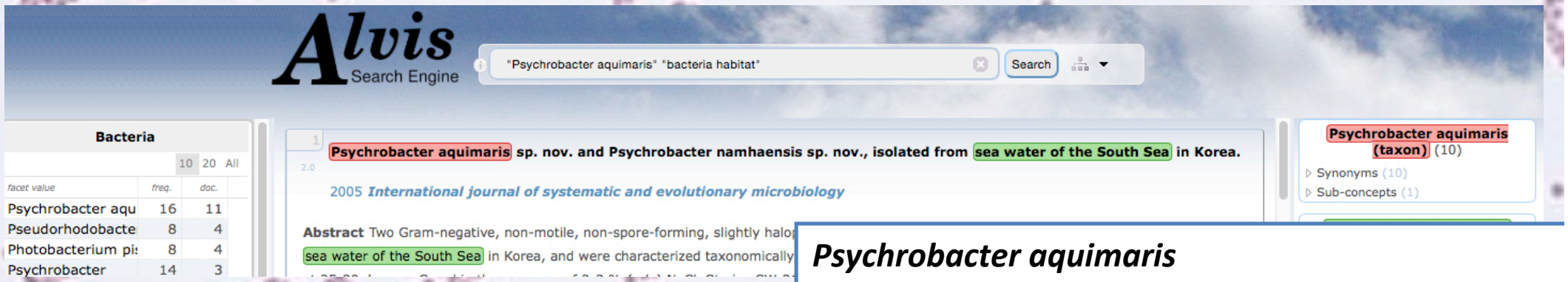
Psychrobacter aquimaris

marine environment

marine environment

Toute l'information de PubMed, analysée et accessible en ligne

- 1,16 millions de documents
- 3,63 millions de relations bactérie - habitat, extraites et normalisées



Psychrobacter aquimaris

Jamais identifiée dans des aliments

Très présente dans les milieux marins

Explication : elle est venue avec le sel ajouté

OntoBiotope, une application TDM pour les microbiologistes

- **Extrait** précisément les information sur les habitats des microorganismes et les lier entre elles.
- Associe l'information à des **catégories hiérarchiques**, taxon, habitat, phénotype, ... pour la "standardiser" à différents niveaux de généralité.
- **Intègre et analyse** les informations sur la plateforme bioinformatique IFB-Migale pour détecter des anomalies et produire automatiquement des hypothèses.

Recherche et ingénierie pluridisciplinaire

Analyse metagénomique, identification des souches

NGS, bioinformatique et bases de données,
assemblage et comparaison de séquences

R & D biologie
R&D bioinformatique
Infrastructures

Conception de l'ontologie OntoBiotope

IA, Traitement Automatique de la Langue,
Apprentissage Automatique, Représentation des Connaissances

Recherche appliquée en informatique
Expertise microbiologie
Expertise aliment

Extraction de l'information à partir de textes, concepts et relations

Sciences de l'Information (corpus)
Traitement Automatique de la Langue, Apprentissage Automatique,
Représentation des Connaissances. Suite logicielle Alvis.

Recherche fondamentale et appliquée en informatique
Ingénierie logicielle

Evaluation par une *shared task* Bacteria Biotope (BioNLP-ST'16)

Recherche appliquée en informatique

Intégration des informations et mise à disposition (Migale)

Bases de données et web sémantique. Workflows.

R & D informatique
R & D bioinformatique
Infrastructures

Extraction d'information, quelle spécificité en biologie ?

Alvis Search Engine

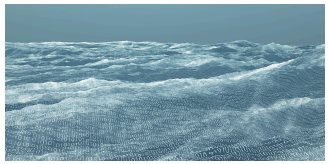
Repose sur un processus automatique de TDM d'une grande complexité, basé sur des outils d'intelligence artificielle

- Prédiction par apprentissage automatique
- Analyse sémantique profonde par traitement automatique de la langue

Configuration

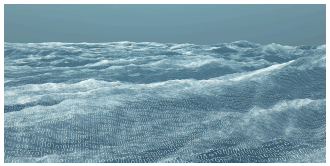
openMINTED

OpenMinTeD, infrastructure européenne de TDM (projet européen H2020)



Enjeu majeur pour l'Europe : mettre à disposition de tous les chercheurs, une plateforme libre de TDM, workflows, composants et ressources.

Adaptation



Apprentissage automatique à partir de données annotées manuellement

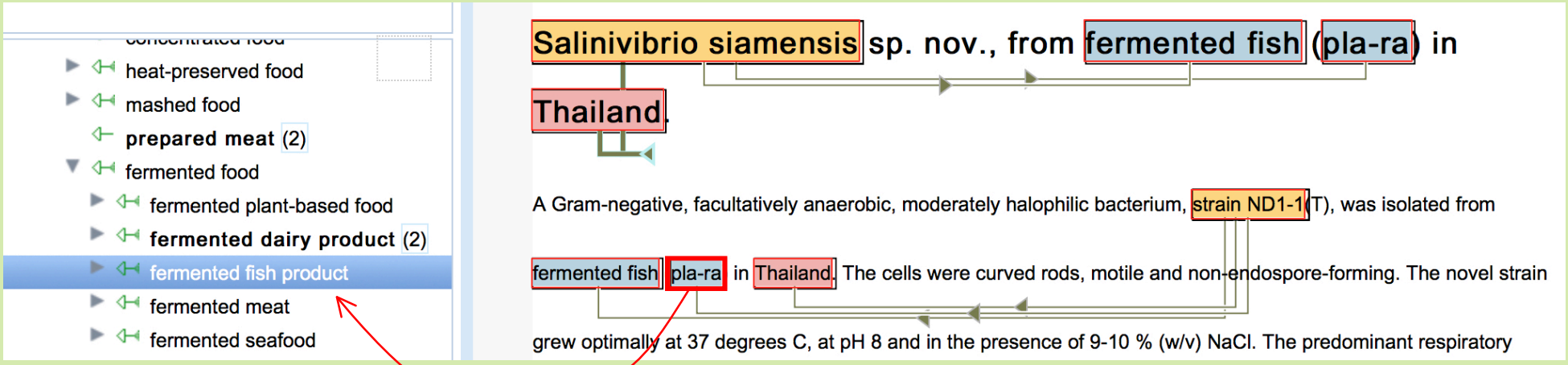
Utilisation de **ressources linguistiques** spécialisées (corpus, lexiques, ontologies)

Exemple de texte annoté pour l'apprentissage

AlvisAE associe une représentation formelle au texte libre

Ontologie

Termes annotés et relations



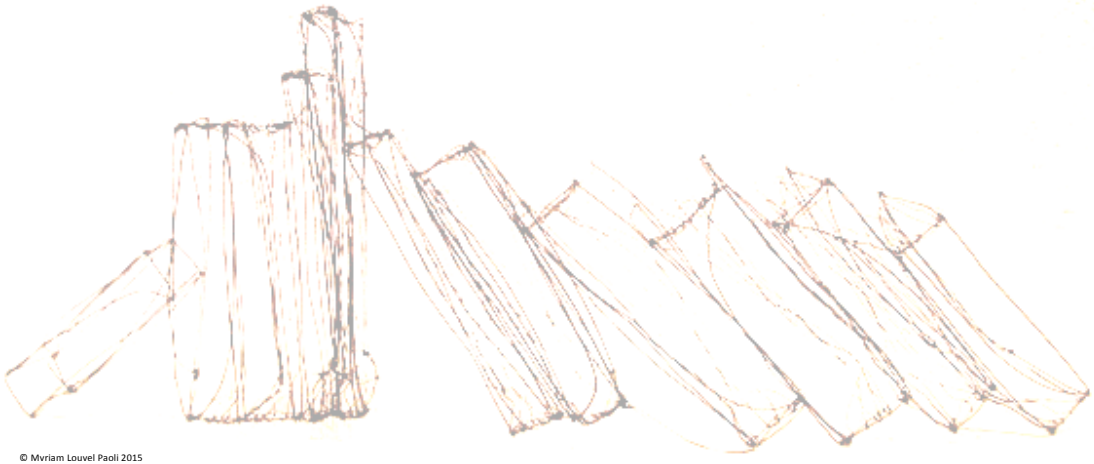
Rattachement du terme au concept

Des bases du dialogue entre informatique et biologie

Fil conducteur

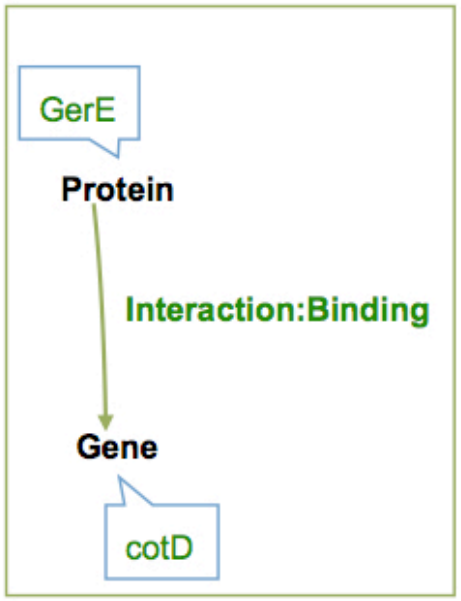
- En biologie : modéliser un réseau de régulation de gènes (collaboration Inra : IJPB et MaIAGE)
- En informatique : extraire automatiquement des informations à partir de textes.
- Faire ensemble, comment ?

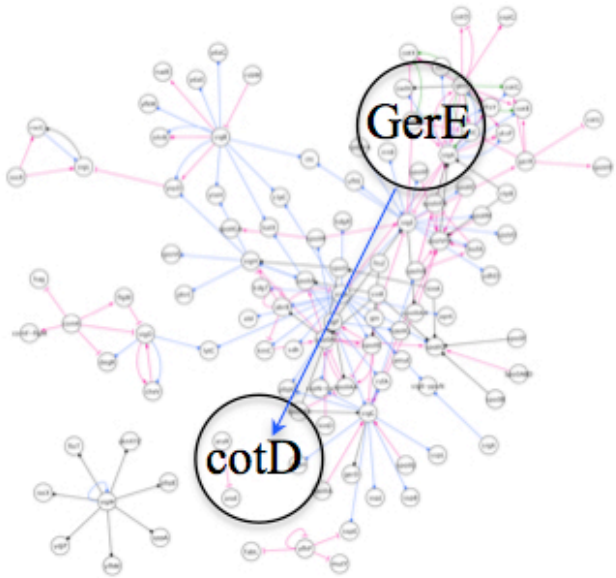
Travailler sur un objet à la fois concret et virtuel
dans un but commun



Objectif, extraction systématique de l'information de régulation

*We show that **GerE** binds to two **sites** that span the -35 region of the **cotD** **promoter***

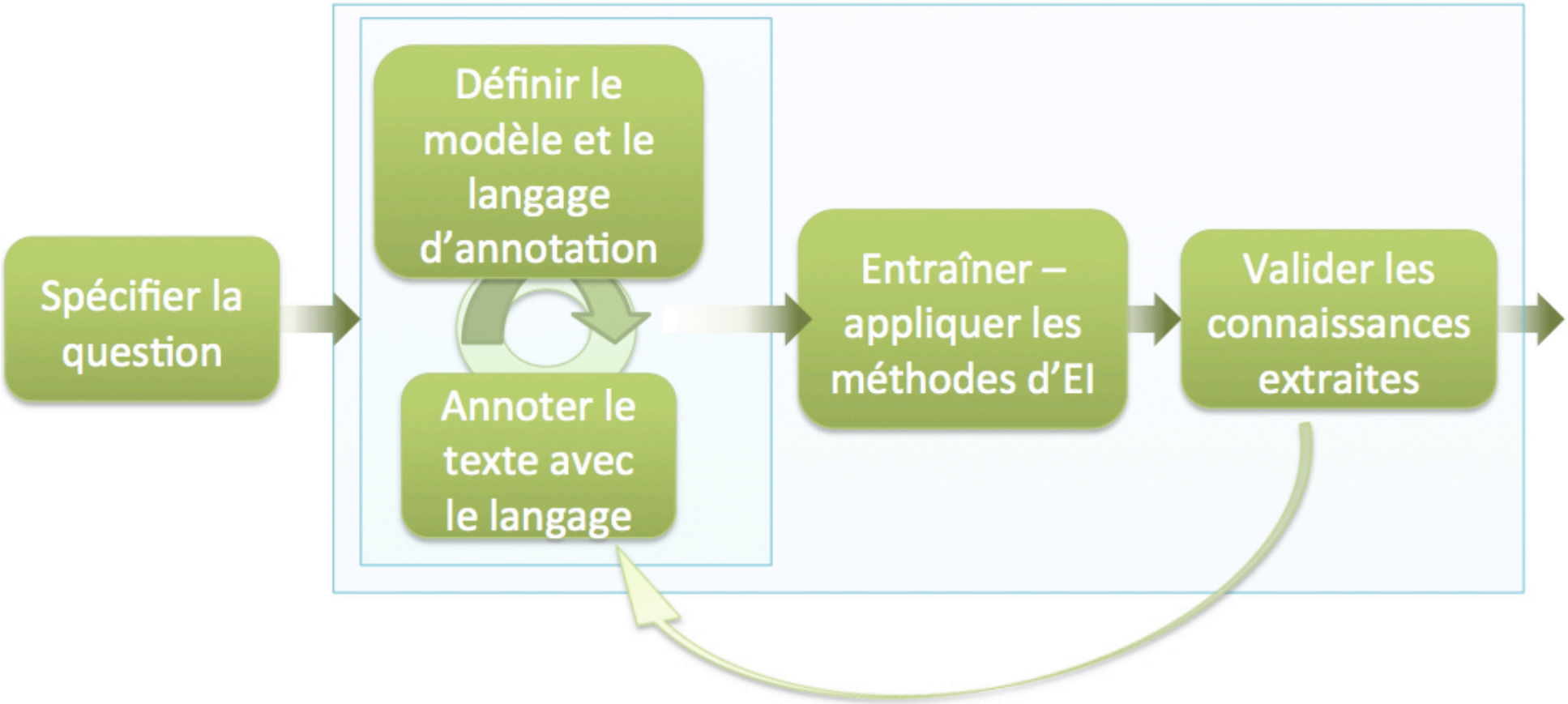


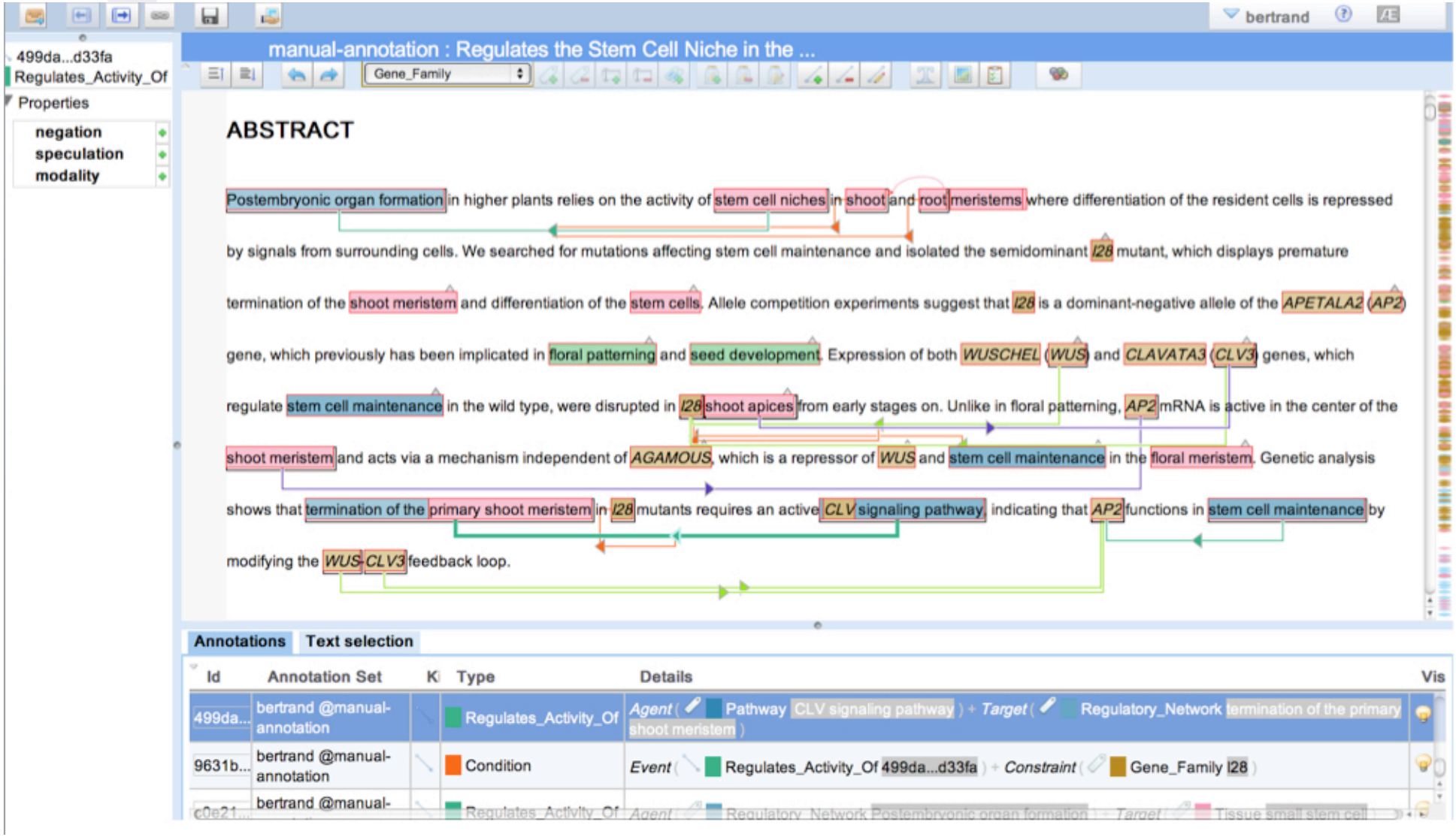


Extraction à partir du texte

Modélisation du réseau de régulation

Principe, itératif





Exemples d'annotation manuelle avec le modèle final

Le modèle : des concepts et des relations

Molecule		DNA	Gene
			Gene_Family
			Box
			Promoter
	DNA Product		RNA
		Amino acid sequence	Protein
			Protein_Family
			Protein_Complex
			Protein_Domain
			Hormone
Dynamic Process			Regulatory_Network
			Metabolic pathway
Context		Biological context	Genotype
			Tissue
			Development_Phase
			Environmental_Factor

Where and When

- Presence_In_Genotype
- Occurrence_In_Genotype
- Presence_At_Stage
- Occurrence_During
- Localization

Function

- Involvement_In_Process
- Transcription_Or_Translation
- Functional_Equivalence

Regulation

- Regulation_Of_Accumulation
- Regulation_Of_Development_Phase
- Regulation_Of_Expression
- Regulation_Of_Molecule_Activity
- Regulation_Of_Process
- Regulation_Of_Tissue_Development

Composition and Membership

- Primary_Structure_Composition
- Protein_Complex_Composition
- Protein_Domain_Composition
- Family_Membership
- Sequence_Identity

Interaction

- Interaction
- Binding

Construction d'un espace partagé interdisciplinaire

Quelle pédagogie d'apprentissage mutuel?

- Compréhension mutuelle à partir d'exemples
- Exigence d'une interprétation commune et non ambiguë apportée par la formalisation
- Annotation par plusieurs biologistes : confrontation des points de vue, analyse des divergences d'interprétation du texte guidée par le modèle
- Généralisation et abstraction par analogie : réinvestissement des acquis dans l'annotation d'autres exemples.
- Mise en commun et explicitation des définitions par la rédaction commune du guide : règles générales et exemples, sorte de *Bible*.
- Démarche itérative : aller-retour entre modèle – exemples ⇒ Réévaluation critique du modèle ⇒ Vers une convergence du texte et du modèle de connaissance.

Qu'est-ce que le biologiste apprend de l'informatique ?

Une figure imposée ou un enrichissement mutuel ?

- Apprendre sur son domaine en confrontant son point de vue, ses pré-supposés à d'autres biologistes et aux informaticiens.
- Remet en question des systèmes de pensée considérés comme acquis ou valide
- Pertinence linguistique de l'annotation -> apprendre à comprendre et à écrire dans la perspective d'une analyse automatique, réduire l'ambiguïté linguistique.
- Homogénéité pour l'apprentissage, exigence du formel. Familiarisation avec la démarche de modélisation.
- Limites de l'expressivité du langage formel : on ne peut pas tout exprimer.
- Annotation du texte et raisonnement après extraction (inférence) : familiarisation avec le raisonnement automatique.
- Apprendre à expliquer, à hiérarchiser les connaissances en fonction de leur importance. Apprendre à expliciter les articulations entre les connaissances (causalité, influence, …). Distinguer le sûr de l'incertain.

Qu'est-ce que l'informaticien apprend de la biologie?

- Objets biologiques et leurs interactions, à différents niveaux, moléculaires, physiologique. Une grande complexité.
- Articulation entre le cadre général biologique – et la question biologique particulière.
- Distinctions fondamentales *versus* distinctions secondaires ou inutiles.
- Accepter les incertitudes, le flou. La biologie, une science en marche. Comment le représenter ?
- Les limites de la modélisation : ce qu'on ne sait pas distinguer ou nommer, des questions pour le futur.
- Renforcer ses compétences particulières en modélisation pour la biologie.

Plus généralement

Recettes de base

- collaboration stratégique pour chaque partie
- publications ET co-publications
- partager un calendrier commun sur une longue durée (risque de décalage)
- ne pas sous-estimer le temps de l'analyse du besoin
- grande curiosité pour le domaine de l'autre
- chasser les incompréhensions, les incohérences
  (ne pas laisser filer, par lassitude).

Et l'enseignement en informatique ?

- Contribution institutionnelle et des personnes à des formations supérieures
  ex. AgroParisTech, écoles vétérinaires
- Contribution de l'Inra à l'élaboration de politique de sites (ex. Université Paris-Saclay)
- Implication au niveau des personnes à la formation supérieure, écoles et universités
- Formation permanente (ex. "Bioinformatique par la pratique", 145 personnes par an)
- Organisation d'écoles chercheurs.