Centre de recherche INRIA Rennes – Bretagne Atlantique

L'INRIA

Le Centre

La bio-informatique

Le bio-informaticien

Pierre Peterlongo
Mai 2011
Lycée Descartes Rennes

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
RENNES - BRETAGNE ATLANTIQUE

Les instituts de recherche en France

Etablissements publics à caractère scientifique et technologique (EPST) CEMAGREF Centre national du machinisme agricole, du génie rural, des eaux et des forêts
CNRS Centre national de la recherche scientifique
INED Institut national d'études démographiques
INRA Institut national de la recherche agronomique
INRETS Institut national de recherche sur les transports et leur sécurité
INRIA Institut national de recherche en informatique et en automatique
INSERM Institut national de la santé et de la recherche médicale
IRD Institut de recherche pour le développement
LCPC Laboratoire central des ponts et chaussées
Etablissements publics à caractère industriel et commercial (EPIC) ADEME Agence de l'environnement et de la maîtrise de l'énergie
ADIT Agence pour la diffusion de l'information technologique
ANDRA Agence nationale de gestion des déchets radioactifs
BRGM Bureau de recherches géologiques et minières
CEA Commissariat à l'énergie atomique
CIRAD Centre de coopération international en recherche agronomique
CNES Centre national d'études spatiales
CSI Cité des sciences et de l'industrie
CSTB Centre scientifique et technique du bâtiment
IFP Institut français du pétrole http://www.ifp.fr
IFREMER Institut français de recherche pour l'exploitation de la mer
INERIS Institut national de l'environnement industriel et des risques
IRSN Institut de radioprotection et de sûreté nucléaire
ONERA Office national d'études et de recherches aérospatiales
OSEO (OSEO Anvar, ex Agence nationale de valorisation de la recherche)
Etablissements publics à caractère administratif (EPA)
CEE Centre d'études de l'emploi
CINES Centre informatique national de l'enseignement supérieur
CNDP Centre national de documentation pédagogique *(Etablissement rattaché à l'Enseignement scolaire)*
CNED Centre national d'enseignement à distance *(Etablissement rattaché à l'Enseignement scolaire)*
CNOUS Centre National des Œuvres Universitaires et Scolaires
GET Groupe des Ecoles des Télécommunications (Recherche en technologie de l'information et de la communication)
INRP Institut national de recherche pédagogique
EPAUnivSav : désamiantage, mise en sécurité et rénovation du site
INRAP Institut national de recherches archéologiques préventives **Fondations**
CEPH Centre d'étude du polymorphisme humain
Institut Curie
Institut Pasteur de Paris
Institut Pasteur de Lille
Groupements d'intérêt public (GIP) ANRS Agence nationale de la recherche sur le sida
CNRGI Consortium national de recherche en génomique (dont fait partie le Centre National de génotypage) avec :
CNSGénoscope - Centre national de séquençage ; RNG Réseau national des génopoles
CNRS-CNRS Centre National de Recherche sur les Sites et Sols Pollués
IPEV Institut polaire français Paul-Emile Victor
GENOPOLE GIP consacré à la recherche en génomique et au développement d'entreprises de biotechnologies
OST Observatoire des sciences et techniques
RENATER Réseau national pour la technologie, l'enseignement et la recherche
Etablissement professionnel
Etablissements d'enseignement supérieur et de recherche 82 Universités (liste complète: http://www.education.gouv.fr/sup/univ.htm)
Instituts nationaux polytechniques
Grenoble, www.inpg.fr
Lorraine Nancy http://www.inpl-nancy.fr
Toulouse-www.univ-inpt.fr
4 Ecoles normales supérieures
Ecole normale supérieure http://www.ens.fr
Ecole normale supérieure de Cachan http://www.ens-cachan.fr
Ecole normale supérieure Lettre et Sciences humaines http://www.ens-lsh.fr
Ecole normale supérieure de Lyon http://www.ens-lyon.fr
5 Ecoles françaises à l'étranger
Casa de Velazquez http://www.casadevelazquez.org
École française de Rome http://www.ecole-francaise.it
École française archéologie athènes http://www.efa.gr
École Française d'Extrême-Orient http://www.efeo.fr
Institut français d'archéologie orientale du Caire IFAO. http://www.ifao.egnet.net/doc/AccIFAO.htm
14 Grande Etablissements de statuts divers
CNAM (Conservatoire National des Arts et Métiers) http://www.cnam.fr
Collège de France http://www.college-de-france.fr
Ecole Centrale des Arts et Manufactures http://www.ecp.fr
Ecole Nationale des Chartes http://www.enc.sorbonne.fr
(ENSAM) Ecole Nationale Supérieure d'Arts et Métiers http://www.ensam.fr
Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques http://www.enssib.fr
(EPHE) Ecole Pratique des Hautes Etudes http://www.ephe.sorbonne.fr
(EHESS) Ecole des Hautes Etudes en Sciences Sociales http://www.ehess.fr
(GET) Groupe des écoles des Télécommunications http://www.get-telecom.fr
(IEP) Institut d'Etudes Politiques de Paris http://www.sciences-po.fr
(INALCO) Institut National des Langues et Civilisation Orientales http://www.inalco.fr/pub
MNHN (Muséum National d'Histoire naturelle) http://www.mnhn.fr
Observatoire de Paris http://www.obspm.fr
Palais de la Découverte http://www.palais-decouverte.fr Groupe des Ecoles des Télécommunications (GET) (établissement public placé sous la tutelle de la ministre déléguée à l'industrie: http://www.get-telecom.fr)

Etablissements publics à caractère scientifique, culturel et professionnel. (EPCSCP)
Ecole centrale de Lille http://www.ec-lille.fr
Ecole centrale de Lyon http://www.ec-lyon.fr/
Ecole centrale de Nantes http://www.ec-nantes.fr
Ecole nationale des Ponts et Chausées http://www.enpc.fr/
Ecole nationale supérieure des arts et industries de Strasbourg http://www-ensais.u-strasbg.fr/
Institut de Physique du Globe de Paris http://www.ipgp.jussieu.fr
Institut national des sciences appliquées de Lyon http://www.insa-lyon.fr/sommaire.html
Institut national des sciences appliques de Rennes http://www.insa-rennes.fr
Institut national des sciences appliquées de Toulouse www.insa-toulouse.fr
Institut national des sciences appliquées de Rouen http://www.insa-rouen.fr/indexsom.html
Institut supérieur des matériaux et de la construction mécanique http://www.cefi.org/fraECOLES/ecole_249.html
Université de technologie de Compiègne http://www.utc.fr
Université de technologie de Belfort-Montbéliard http://www.utbm.fr
Université de technologie de Troyes http://www.univ-troyes.fr/sommaire.htm

http://www2.enseignementsup-recherche.gouv.fr

Les instituts de recherche en France

- CNRS: un peu tout
- INED: Démographie
- IRD: Développement
- CNES: Etudes spaciales
- IFREMER: Mer
- INRA: Agronomie
- **INRIA: Informatique**
- 82 Universités...

Etablissements publics à caractère scientifique et technologique (EPST) CEMAGREF Centre national du machinisme agricole, du génie rural, des eaux et des forêts
CNRS Centre national de la recherche scientifique
INED Institut national d'études démographiques
INRA Institut national de la recherche agronomique
INRETS Institut national de recherche sur les transports et leur sécurité
INRIA Institut national de recherche en informatique et en automatique
INSERM Institut national de la santé et de la recherche médicale
IRD Institut de recherche pour le développement
LCPC Laboratoire central des ponts et chaussées
Etablissements publics à caractère industriel et commercial (EPIC) ADEME Agence de l'environnement et de la maîtrise de l'énergie
ADIT Agence pour la diffusion de l'information technologique
ANDRA Agence nationale de gestion des déchets radioactifs
BRGM Bureau de recherches géologiques et minières
CEA Commissariat à l'énergie atomique
CIRAD Centre de coopération international en recherche agronomique
CNES Centre national d'études spatiales
CSI Cité des sciences et de l'industrie
CSTB Centre scientifique et technique du bâtiment
IFP Institut français du pétrole http://www.ifp.fr
IFREMER Institut français de recherche pour l'exploitation de la mer
INERIS Institut national de l'environnement industriel et des risques
IRSN Institut de radioprotection et de sûreté nucléaire
ONERA Office national d'études et de recherches aérospatiales
OSEO (OSEO Anvar, ex Agence nationale de valorisation de la recherche)
Etablissements publics à caractère administratif (EPA)
CEE Centre d'études de l'emploi
CINES Centre informatique national de l'enseignement supérieur
CNDP Centre national de documentation pédagogique *(Etablissement rattaché à l'Enseignement scolaire)*
CNED Centre national d'enseignement à distance *(Etablissement rattaché à l'Enseignement scolaire)*
CNOUS Centre National des Œuvres Universitaires et Scolaires
GET Groupe des Ecoles des Télécommunications (Recherche en technologie de l'information et de la communication)
INRP Institut national de recherche pédagogique
EPA Jussieu : désamiantage, mise en sécurité et rénovation du site
INRAP Institut national de recherches archéologiques préventives **Fondations**
CEPH Centre d'étude du polymorphisme humain
Institut Curie
Institut Pasteur de Paris
Institut Pasteur de Lille
Groupements d'intérêt public (GIP) ANRS Agence nationale de la recherche sur le sida
CNRGI Consortium national de recherche en génomique (dont fait partie le Centre National de génotypage) avec :
CNS Génoscope - Centre national de séquençage ; RNG Réseau national des génopoles
CNRS-SO Centre National de Recherche sur les Sites et Sols Pollués
IPEV Institut polaire français Paul-Emile Victor
GENOPOLE GIP consacré à la recherche en génomique et au développement d'entreprises de biotechnologies
OST Observatoire des sciences et techniques
RENATER Réseau national pour la technologie, l'enseignement et la recherche
Etablissement professionnel
Etablissements d'enseignement supérieur et de recherche 82 Universités (liste complète: http://www.education.gouv.fr/sup/univ.htm)
Instituts nationaux polytechniques
Grenoble, www.inpg.fr
Lorraine Nancy http://www.inpl-nancy.fr
Toulouse-www.univ-inpt.fr
4 Ecoles normales supérieures
Ecole normale supérieure http://www.ens.fr
Ecole normale supérieure de Cachan http://www.ens-cachan.fr
Ecole normale supérieure Lettre et Sciences humaines http://www.ens-lsh.fr
Ecole normale supérieure de Lyon http://www.ens-lyon.fr
5 Ecoles françaises à l'étranger
Casa de Velazquez http://www.casadevelazquez.org
École française de Rome http://www.ecole-francaise.it
École française archéologie athènes http://www.efa.gr
École Française d'Extrême-Orient http://www.efeo.fr
Institut français d'archéologie orientale du Caire IFAO. http://www.ifao.egnet.net/doc/AccIFAO.htm
14 Grande Etablissements de statuts divers
CNAM (Conservatoire National des Arts et Métiers) http://www.cnam.fr
Collège de France http://www.college-de-france.fr
Ecole Centrale des Arts et Manufactures http://www.ecp.fr
Ecole Nationale des Chartes http://www.enc.sorbonne.fr
(ENSAM) Ecole Nationale Supérieure d'Arts et Métiers http://www.ensam.fr
Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques http://www.enssib.fr
(EPHE) Ecole Pratique des Hautes Etudes http://www.ephe.sorbonne.fr
(EHESS) Ecole des Hautes Etudes en Sciences Sociales http://www.ehess.fr
(GET) Groupe des écoles des Télécommunications http://www.get-telecom.fr
(IEP) Institut d'Etudes Politiques de Paris http://www.sciences-po.fr
(INALCO) Institut National des Langues et Civilisation Orientales http://www.inalco.fr/pub
MNHN (Muséum National d'Histoire naturelle) http://www.mnhn.fr
Observatoire de Paris http://www.obspm.fr
Palais de la Découverte http://www.palais-decouverte.fr Groupe des Ecoles des Télécommunications (GET) (établissement public placé sous la tutelle de la ministre déléguée à l'industrie: http://www.get-telecom.fr)

Etablissements publics à caractère scientifique, culturel et professionnel. (EPCSCP)
Ecole centrale de Lille http://www.ec-lille.fr
Ecole centrale de Lyon http://www.ec-lyon.fr/
Ecole centrale de Nantes http://www.ec-nantes.fr
Ecole nationale des Ponts et Chausées http://www.enpc.fr/
Ecole nationale supérieure des arts et industries de Strasbourg http://www-ensais.u-strasbg.fr/
Institut de Physique du Globe de Paris http://www.ipgp.jussieu.fr
Institut national des sciences appliquées de Lyon http://www.insa-lyon.fr/sommaire.html
Institut national des sciences appliques de Rennes http://www.insa-rennes.fr
Institut national des sciences appliquées de Toulouse www.insa-toulouse.fr
Institut national des sciences appliquées de Rouen http://www.insa-rouen.fr/indexsom.html
Institut supérieur des matériaux et de la construction mécanique http://www.cefi.org/fraECOLES/ecole_249.html
Université de technologie de Compiègne http://www.utc.fr
Université de technologie de Belfort-Montbéliard http://www.utbm.fr
Université de technologie de Troyes http://www.univ-troyes.fr/sommaire.htm

http://www2.enseignementsup-recherche.gouv.fr

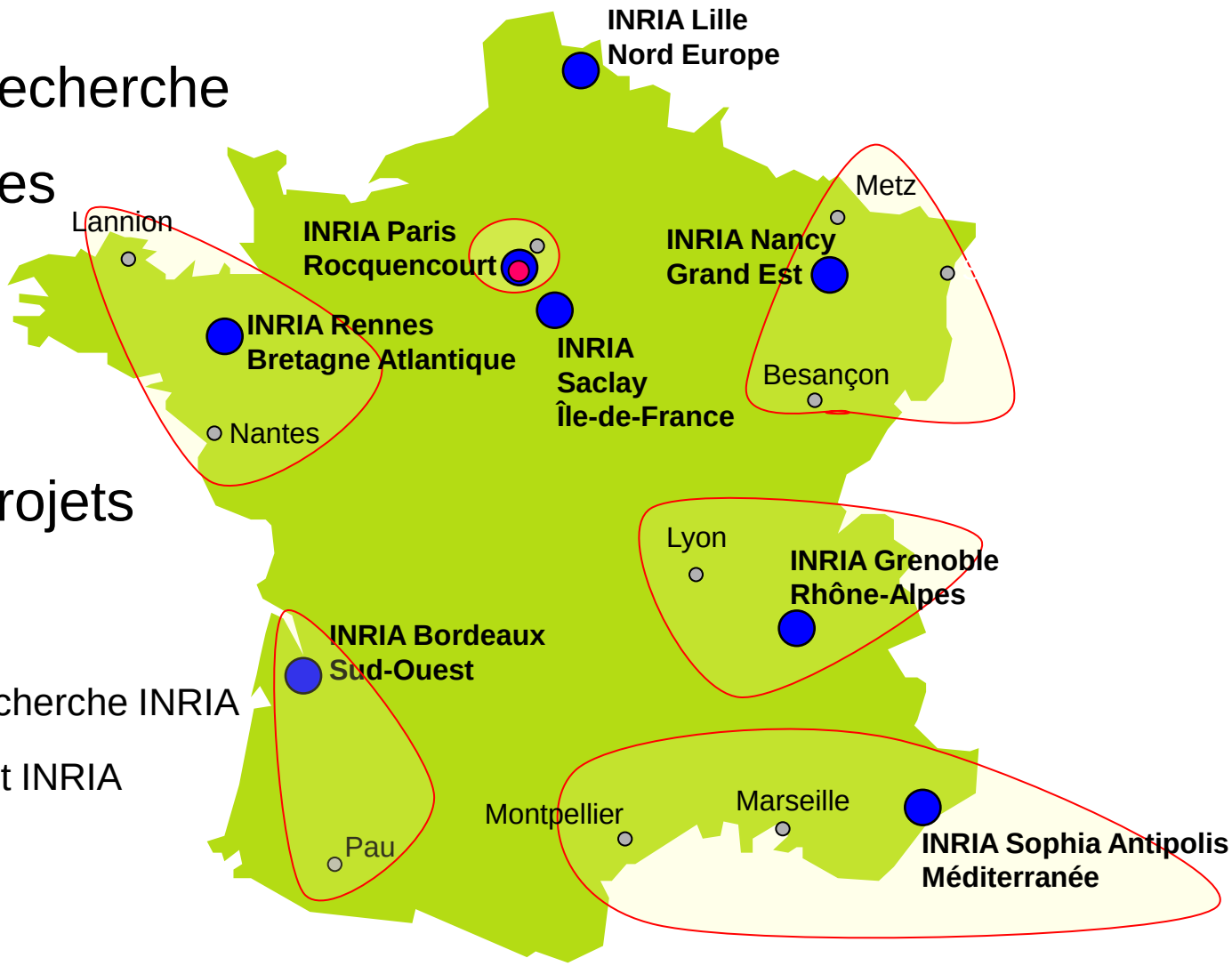
L'INRIA :

8 centres de recherche

3800 personnes

150 équipes-projets

- Siège
- Centre de recherche INRIA
- Équipe-projet INRIA hors site

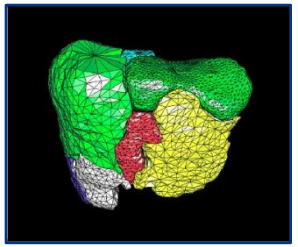


Chiffres : mai 2008

L'INRIA

Ministère de la recherche et **ministère de l'industrie**

créer la référence



expérimenter

chercher

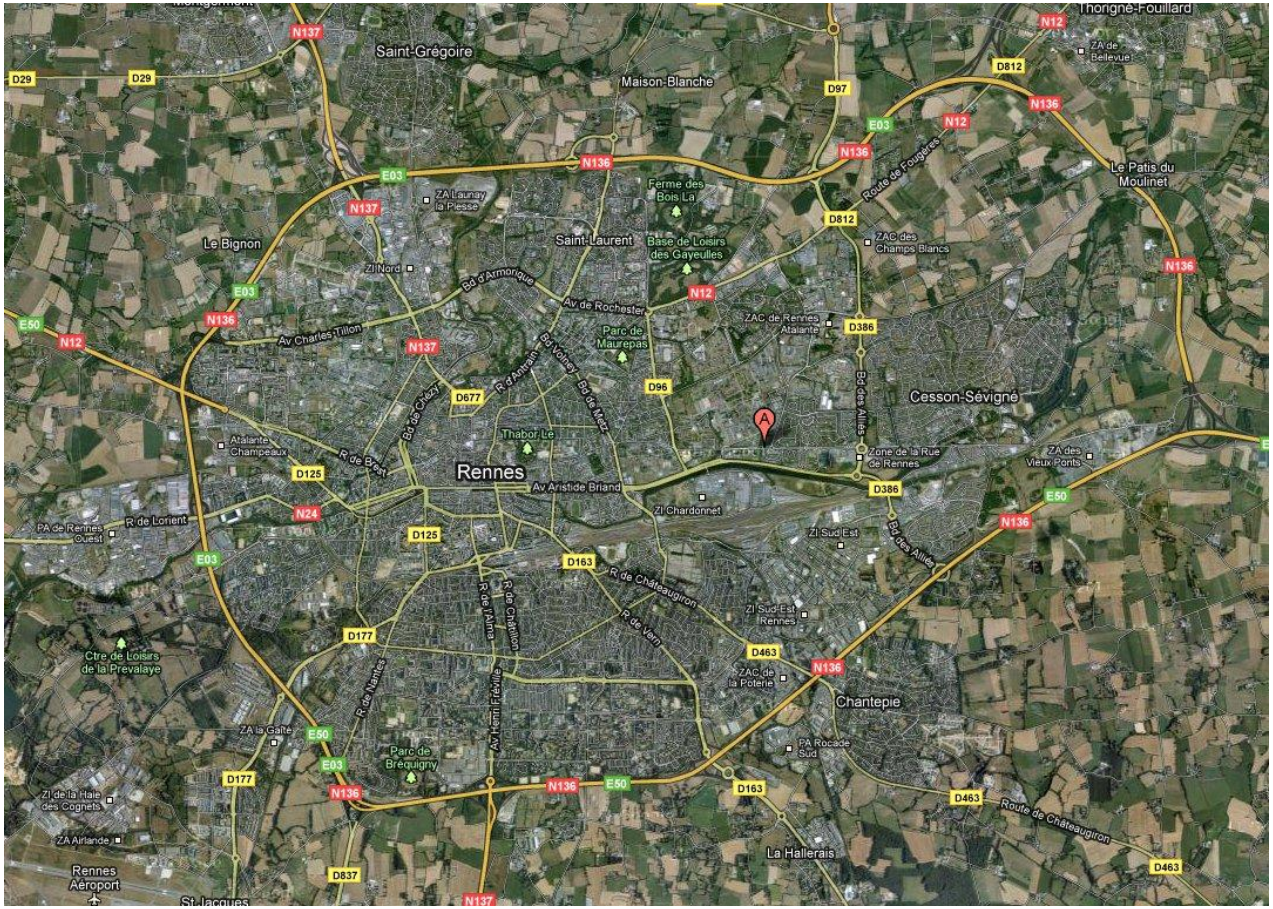
partager & transmettre

Le centre de recherche
INRIA Rennes – Bretagne Atlantique

Présentation générale

Où ?

- **Campus de beaulieu, Rennes.**



Où ?

- Campus de beaulieu, Rennes.

Qui ?

Environ 600 personnes:

- 200 (235) Chercheurs et enseignants chercheurs
- 100 (75) Ingénieurs, techniciens et administratifs

Quoi ?

Sciences du **numérique**

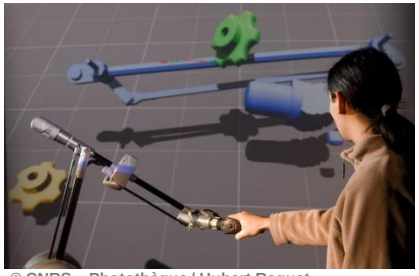
32 équipes de recherche

- Réalité virtuelle
- Réseaux / internet
- Robotique
- Sciences du vivant
- ...

Les plate-formes d'expérimentation

Multimédia

- Système d'immersion virtuelle
- Plate-forme de capture, stockage et indexation de données vidéo



Robotique

- Robots avec capteurs de vision, véhicule Cycab
- Système robotisé d'échographie 3D

Informatique

- Un des 9 nœuds de la grille de calcul nationale Grid'5000
- Plate-forme bio-informatique de Ouest-génopole

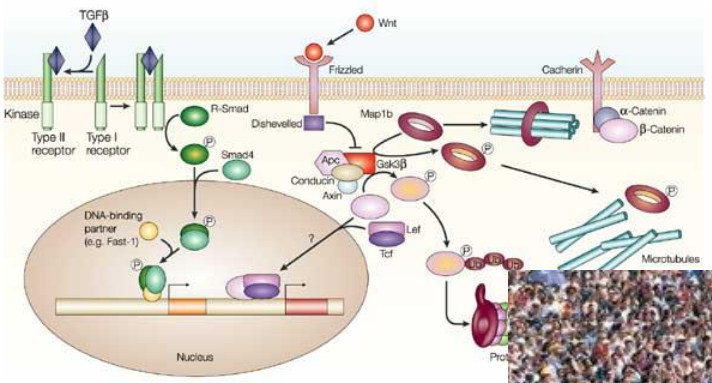
Biomédical

- Plate-forme Neurinfo

L’équipe projet Symbiose

Symbiose, 3 axes de recherche

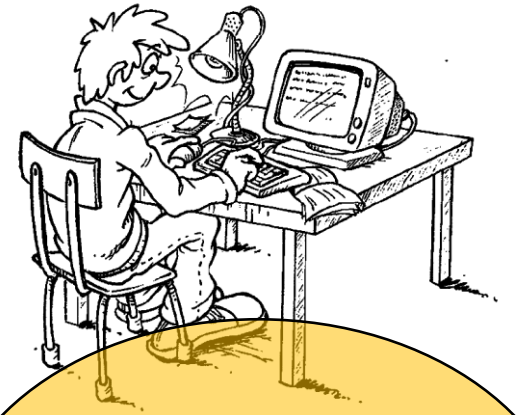
- Interactions



- Grosses masses de données

- Calculs intensifs

Symbiose:
la recherche, la plateforme, les biologistes



Laboratoires de Bio

Centre de ressources

Genouest

Recherche en Bioinfo

Symbiose

Analyses, Projets

Développements

Transferts

C'est quoi la Bioinformatique ?

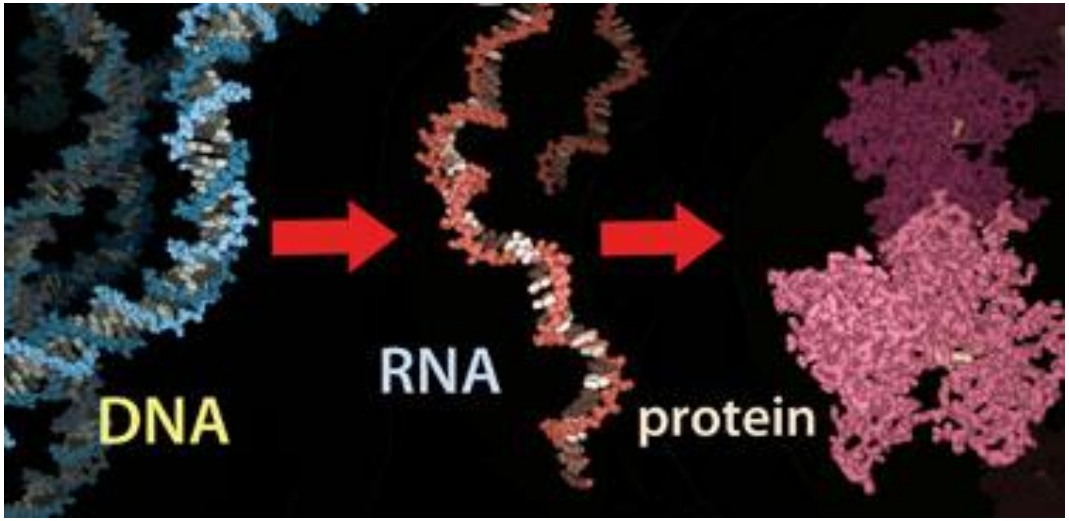
Quand ?

- Naissance : années 80 – 90
- Avant : expérimentations
  - macro ou micro scopique
  - Comportements

Quand ?

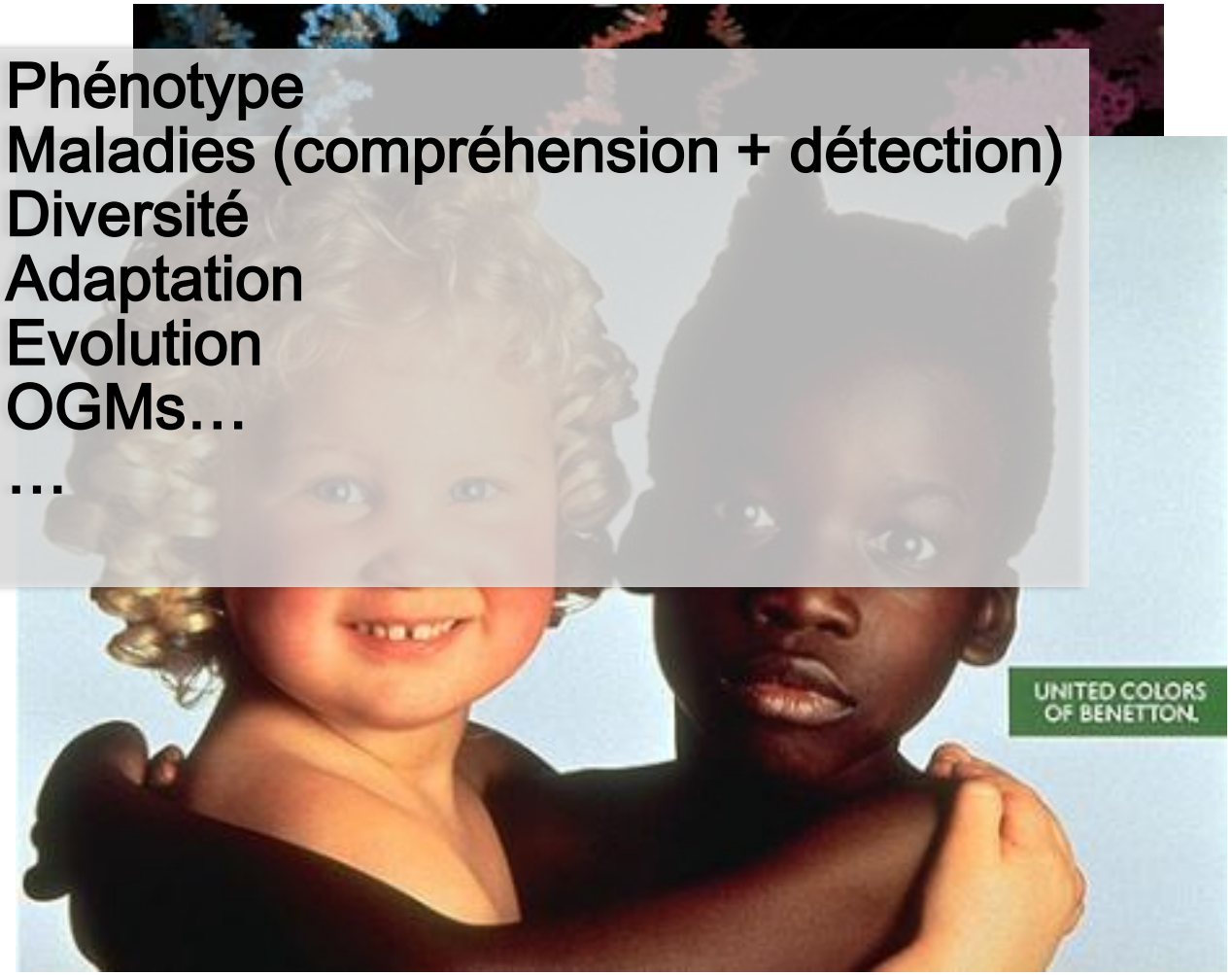
- Naissance : années 80 – 90
- Après : comme avant +
  - ADN, ARN, Acides Aminés...

L'ADN, et après ?



L’ADN, et après ?

L'ADN, et après ?

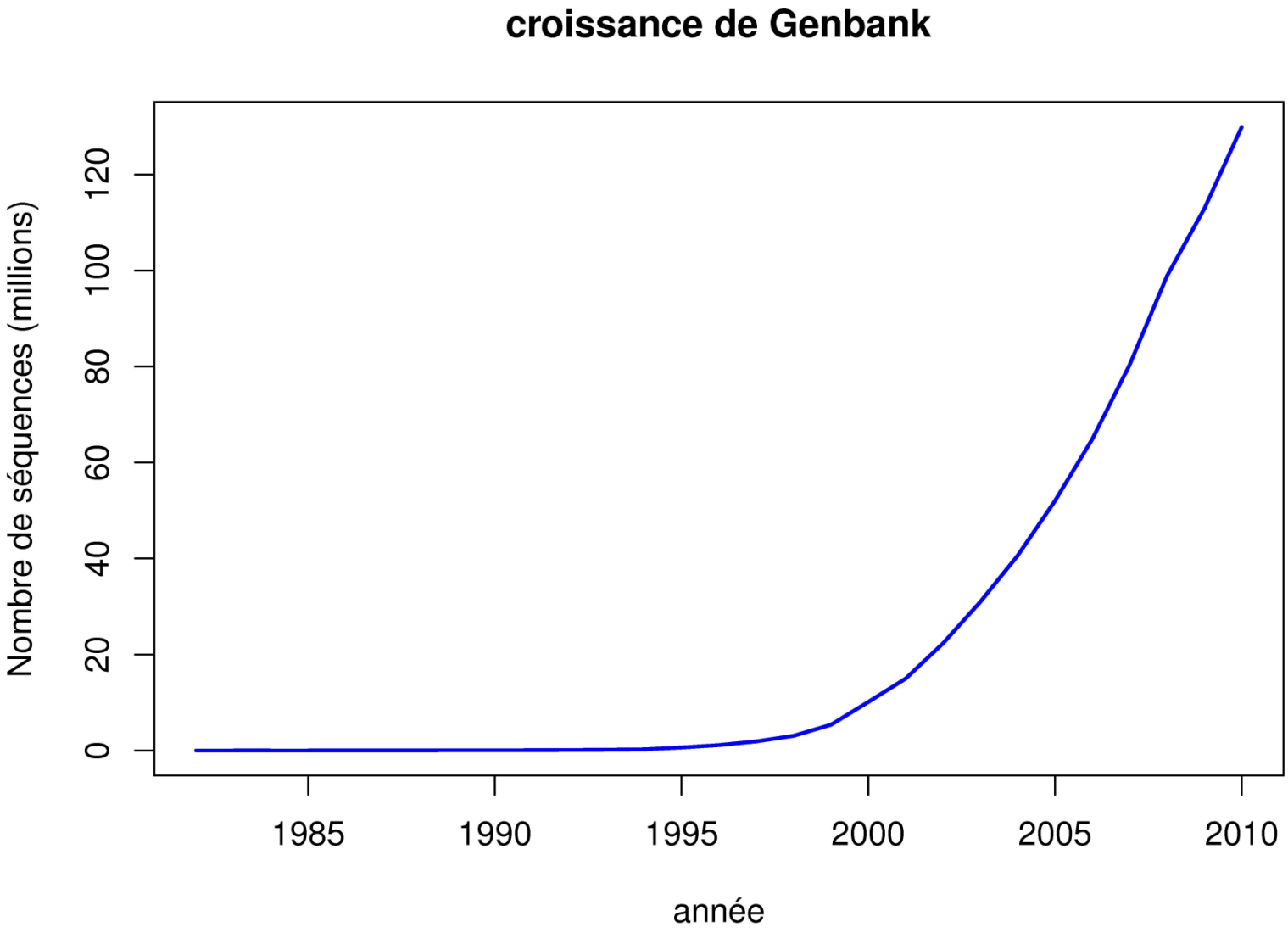


Phénotype
Maladies (compréhension + détection)
Diversité
Adaptation
Evolution
OGMs…
…

Lire l'ADN ?

- 1953 Découverte de la structure en double hélice de l'ADN
- 1977 Développement des techniques de séquençage d'ADN
- 1977 Premier génome séquencé : virus (5 386 nt)
- 1984 Génome du virus du sida (170 000 nt)
- 1990 Début du projet Génome humain
- 1996 Premier génome eucaryote (levure 12 000 000 nt)
- 2001 Premier génome de mammifère (homme 3 000 000 000 nt)
- 2011 >2600 génomes complets (200 mammifères)

Lire l'ADN ?


croissance de Genbank
Nombre de séquences (millions)
0
20
40
60
80
100
120
1985
1990
1995
2000
2005
2010
année

Lire l'ADN = Séquencer

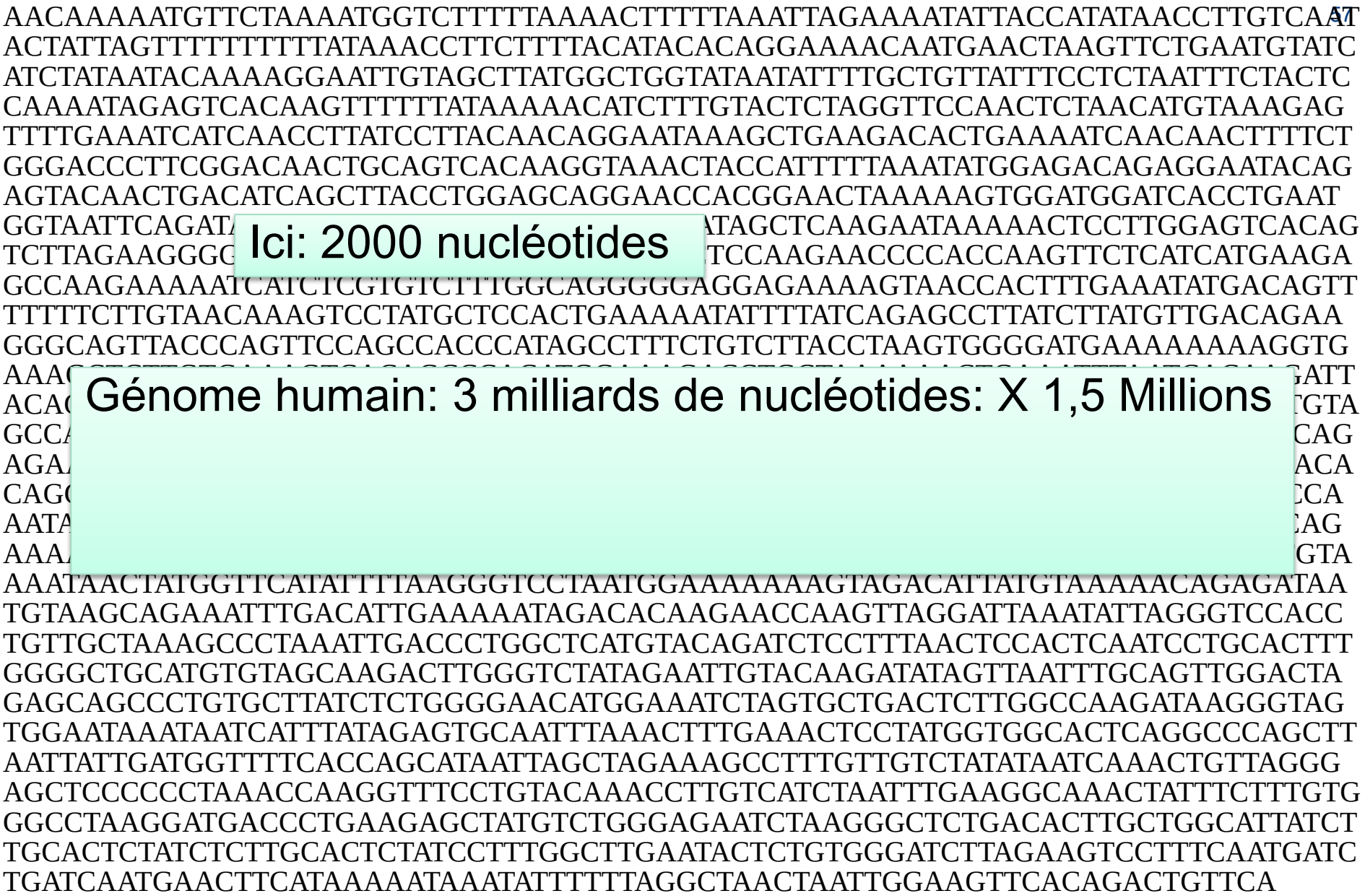
- Techno les plus connues:
  - 454: Genome Sequencer
  - Illumina Solexa Genome Analyser
  - Applied Biosystem: SOLiD

Lire l'ADN, pour quel prix ?

- Génome human:
- 2000: 3 milliards $

- Actuellement: 10000 $

AACAAAAATGTTCTAAAATGGTCTTTTTAAAACTTTTTAAATTAGAAAATATTACCATATAACCTTGTCAAT
ACTATTAGTTTTTTTTTTATAAACCTTCTTTTACATACACAGGAAAACAATGAACTAAGTTCTGAATGTATC
ATCTATAATACAAAAGGAATTGTAGCTTATGGCTGGTATAATATTTTGCTGTTATTTCCTCTAATTTCTACTC
CAAAATAGAGTCACAAGTTTTTTATAAAAACATCTTTGTACTCTAGGTTCCAACTCTAACATGTAAAGAG
TTTTGAAATCATCAACCTTATCCTTACAACAGGAATAAAGCTGAAGACACTGAAAATCAACAACTTTTCT
GGGACCCTTCGGACAACTGCAGTCACAAGGTAAACTACCATTTTTAAATATGGAGACAGAGGAATACAG
AGTACAACTGACATCAGCTTACCTGGAGCAGGAACCACGGAACTAAAAAGTGGATGGATCACCTGAAT
GGTAATTCAGATAAATTGCTGTAAGTTGAGTGTGACATAGCTCAAGAATAAAAACTCCTTGGAGTCACAG
TCTTAGAAGGGGTTCTGCATATTCAGAGGTTTTTACCTCCAAGAACCCCACCAAGTTCTCATCATGAAGA
GCCAAGAAAAATCATCTCGTGTCTTTGGCAGGGGGAGGAGAAAAGTAACCACTTTGAAATATGACAGTT
TTTTTCTTGTAACAAAGTCCTATGCTCCACTGAAAAATATTTTATCAGAGCCTTATCTTATGTTGACAGAA
GGGCAGTTACCCAGTTCCAGCCACCCATAGCCTTTCTGTCTTACCTAAGTGGGGATGAAAAAAAAGGTG
AAAGCTCTTGTGAAAGTCACAGCCCACATGGAAAGACCTGCTAAAAAACTGAAATTTAATCACAAGATT
ACAGAATTCTCTTTTCTCCCATTCCTTGCTACCAAATCAATGGGCATCTAGTATAATAACAGTGGATTTGTA
GCCAAAAGATCTGCAAACTTAGACTCTATTTAAGAATGAGCTCTTAGGGAAACCTAAAGACAACACCAG
AGAAAAAAACAAGGACATTAAAAGAATTGAAAGACTCTGACACATACAGCTACAGAAAACATTGGACA
CAGCCCAAATCCTAGCCAGGTTAAGATACAACCTCACACTAAAGGCCTATGTGTCTCATTCTCTTTTCCA
AATATATGCATATAGCTTTCAATAACAACAAAAAAACTCAAGGTATGCTAAAAAGTAGGAAAAACACAG
AAAATACAAAGCAAGCATTAAACCAAGACTCAGATTTGACAGAAATGTTGGACTTAGAGATGGAATGTA
AAATAACTATGGTTCATATTTTAAGGGTCCTAATGGAAAAAAAGTAGACATTATGTAAAAACAGAGATAA
TGTAAGCAGAAATTTGACATTGAAAAATAGACACAAGAACCAAGTTAGGATTAAATATTAGGGTCCACC
TGTTGCTAAAGCCCTAAATTGACCCTGGCTCATGTACAGATCTCCTTTAACTCCACTCAATCCTGCACTTT
GGGGCTGCATGTGTAGCAAGACTTGGGTCTATAGAATTGTACAAGATATAGTTAATTTGCAGTTGGACTA
GAGCAGCCCTGTGCTTATCTCTGGGGAACATGGAAATCTAGTGCTGACTCTTGGCCAAGATAAGGGTAG
TGGAATAAATAATCATTTATAGAGTGCAATTTAAACTTTGAAACTCCTATGGTGGCACTCAGGCCCAGCTT
AATTATTGATGGTTTTCACCAGCATAATTAGCTAGAAAGCCTTTGTTGTCTATATAATCAAACTGTTAGGG
AGCTCCCCCCTAAACCAAGGTTTCCTGTACAAACCTTGTCATCTAATTTGAAGGCAAACTATTTCTTTGTG
GGCCTAAGGATGACCCTGAAGAGCTATGTCTGGGAGAATCTAAGGGCTCTGACACTTGCTGGCATTATCT
TGCACTCTATCTCTTGCACTCTATCCTTTGGCTTGAATACTCTGTGGGATCTTAGAAGTCCTTTCAATGATC
TGATCAATGAACTTCATAAAAATAAATATTTTTTAGGCTAACTAATTGGAAGTTCACAGACTGTTCA


Ici: 2000 nucléotides
Génome humain: 3 milliards de nucléotides: X 1,5 Millions

Ici: 2000 nucléotides

Génome humain: 3 milliards de nuc s

* ~ 200 livres de 500 pages…

Ici: 2000 nucléotides

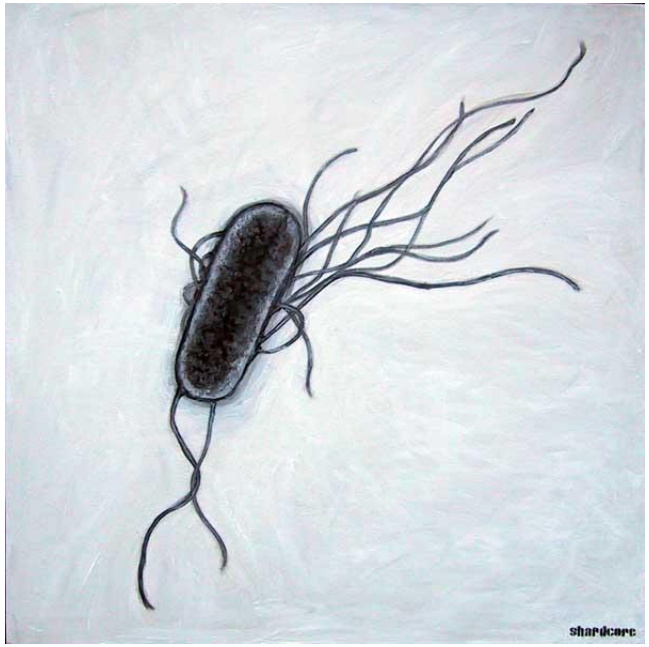
Génome humain: 3 milliards

* ~ 200 livres de

Que faire de tout ça ???
Comment extraire de l'information ?

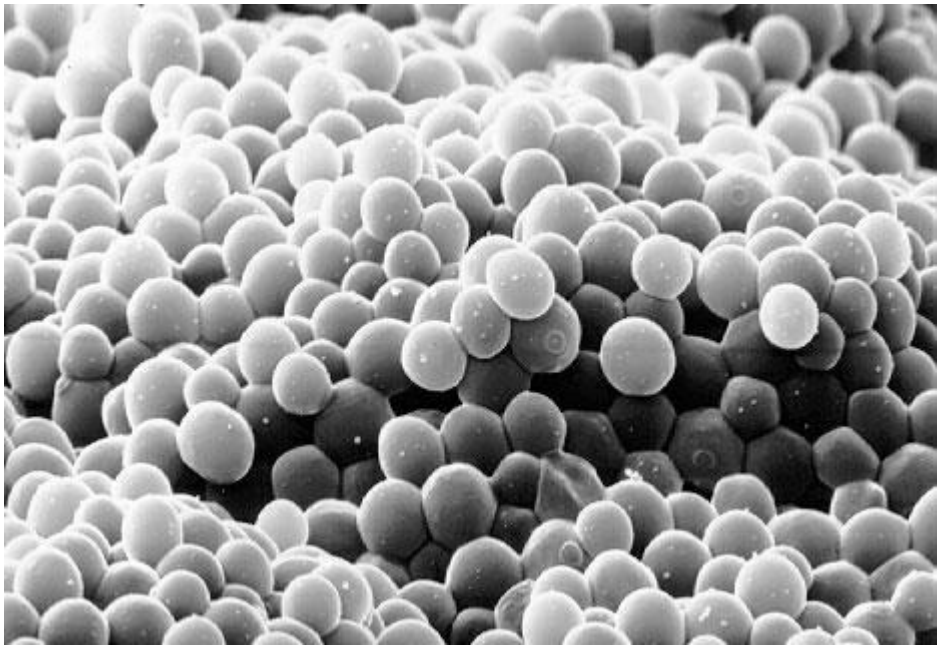
Et on n’est pas seuls…

- E. Coli: 5 Mb



Et on n’est pas seuls…

- E. Coli: 5 Mb
- Levure: 12 Mb

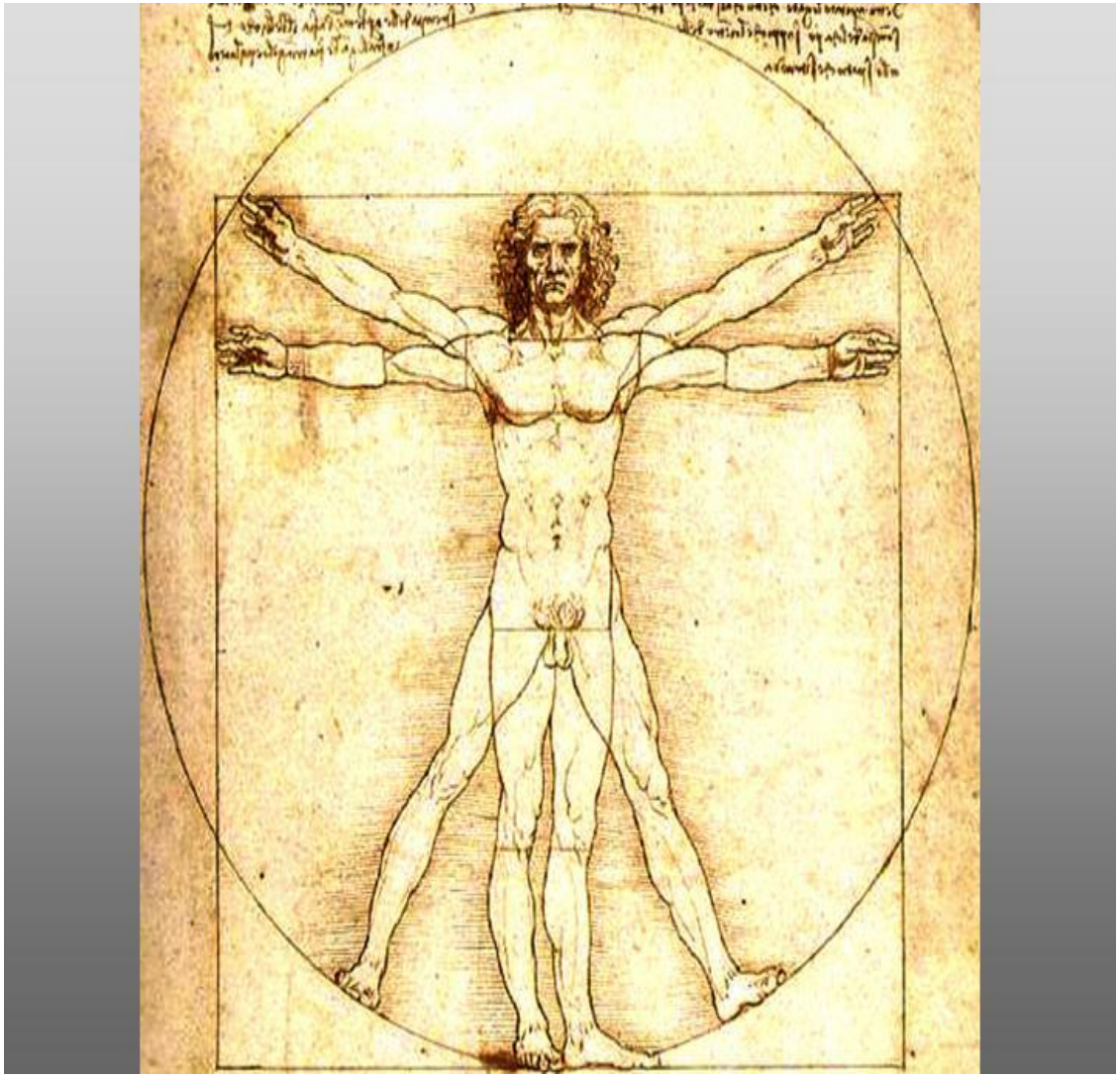


Et on n’est pas seuls…

- E. Coli: 5 Mb
- Levure: 12 Mb
- Mouche: 150 Mb

Et on n'est pas seuls…

- E. Coli: 5 Mb
- Levure: 12 Mb
- Mouche: 150 Mb
- Homme: 3000 Mb



Et on n’est pas seuls…

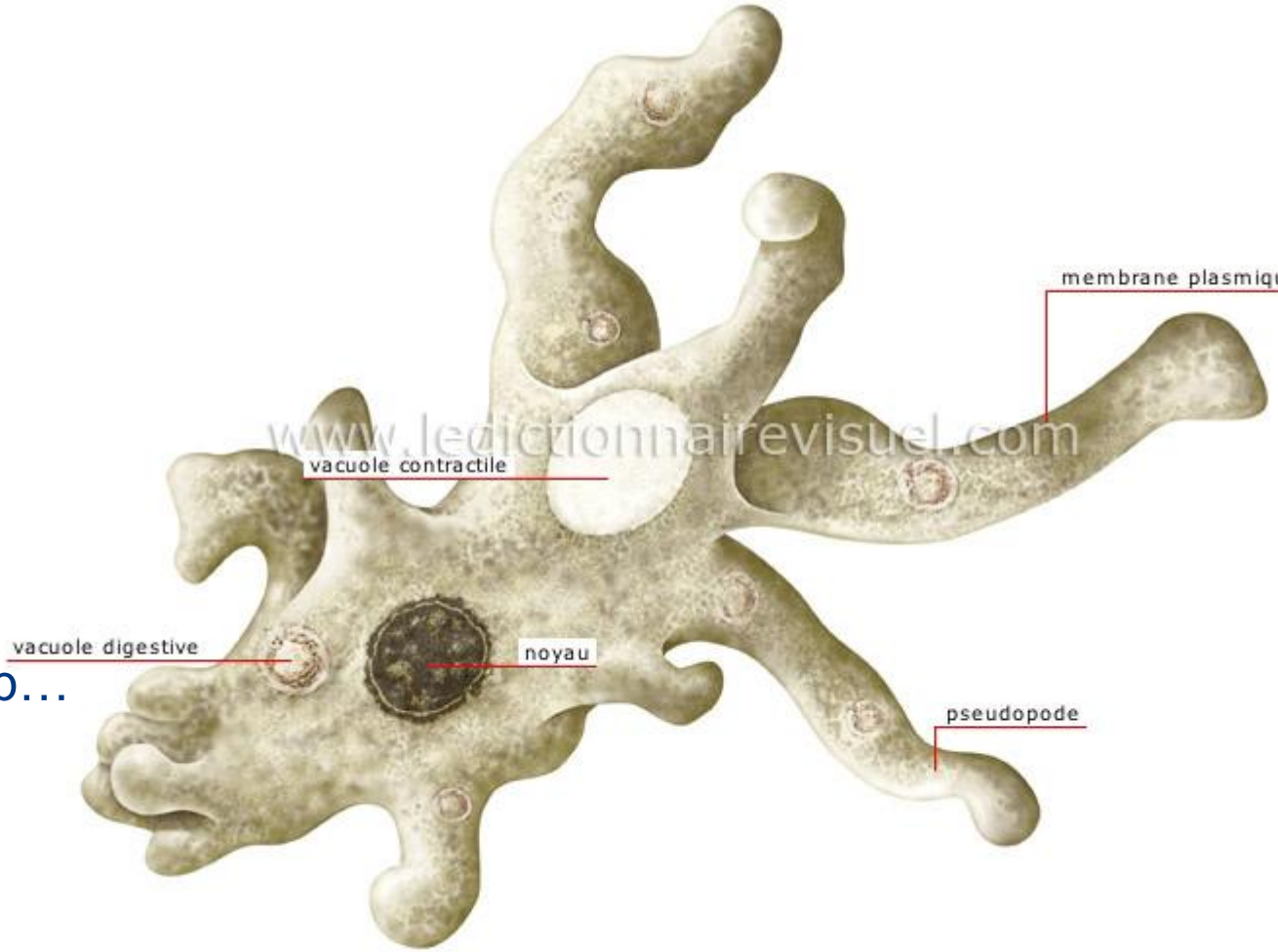
- E. Coli: 5 Mb
- Levure: 12 Mb
- Mouche: 150 Mb
- Homme: 3000 Mb
- Souris: 3400 Mb

Et on n’est pas seuls…

- E. Coli: 5 Mb
- Levure: 12 Mb
- Mouche: 150 Mb
- Homme: 3000 Mb
- Souris: 3400 Mb
- Maïs: 5000 Mb

Et on n’est pas seuls…

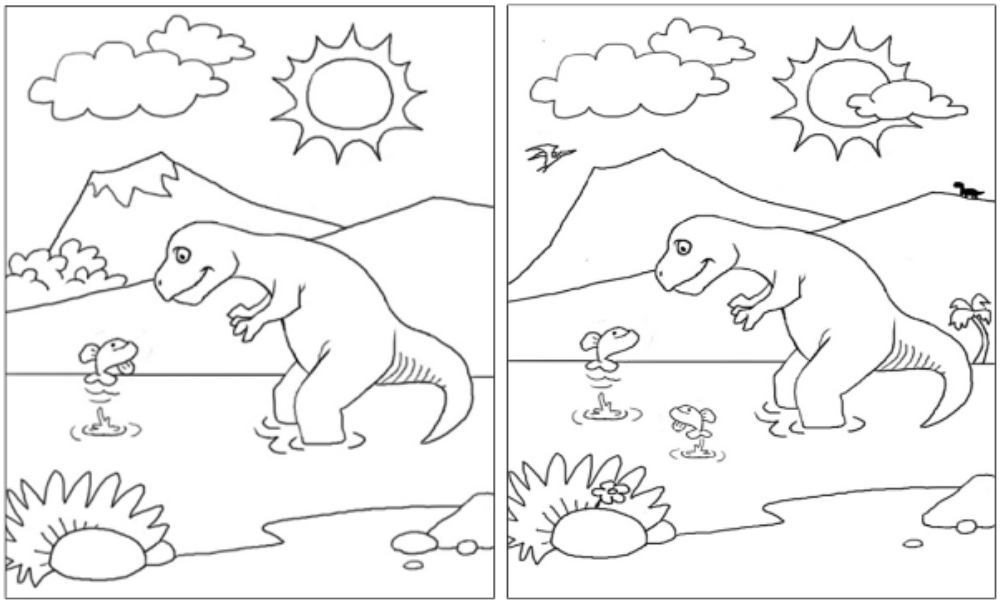
- E. Coli: 5 Mb
- Levure: 12 Mb
- Mouche: 150 Mb
- Homme: 3000 Mb
- Souris: 3400 Mb
- Maïs: 5000 Mb
- Amibe: 675000 Mb…



Que faire de ces données?

→ Les comparer !

- Apprendre en comparant les séquences:
  - Commun = conservé par l'évolution = fonctionnel
  - Différent = "inutile" ou différences en espèces



Que faire de ces données?

→ Les comparer !

- Apprendre en comparant les séquences:
  - Commun = conservé par l'évolution = fonctionnel
  - Différent = "inutile" ou différences en espèces
- Mais aussi :
  - Ne pas recommencer: inférer les connaissances
  - Relations de parenté, mécanismes d'évolution
  - …

Similarité de séquences

- TGATGCCA = TGATGCCA
- TGATGCCA ≈ TGAAGCCA
- TGATGCCA ≈ TGAGCCA

Similarité de séquences (Level 1)

- ATGCTAAATGGGGAAAAGCTGGTACCATACG
- ACGCATGAACGATAAATGGGGAAAATAGGAGT

  - Deux fragments identiques de taille >5 ?

Similarité de séquences (Level 1)

- ATGC**TAAATGGGGAAAA**GCTGGTACCATACG
- ACGCATGAACGA**TAAATGGGGAAAA**TAGGAGT

```
TAAATGGGGAAAA
||||||||||||||
TAAATGGGGAAAA
```

Similarité de séquences (Level 2)

- ATGCTAAATGGGGAAAAGCTGGTACCATACG
- ACAACAGCTGTACTACGCGAGCAGACTACTC

- Deux fragments similaires (85 % identité minimum) de taille >5 ?

Similarité de séquences (Level 2)

- ATGCTAAATGGGG**AAAAGCTGGTAC**CATACG
- AC**AACAGCTGTAC**TACGCGAGCAGACTACTC

```
AAAAGCTGGTAC
||.|||||  |||
AACAGCTG-TAC
```

Similarité de séquences (Level 3)

- ATGCATCGCTAACGCATCAGATGGCTGACTGACTGCATTTTCTCT
ACGATCGGCATCGATCTCAGCATCAGCATCATGCAGCATCATCTA
TACGACCAGCATACTCAGCAGCAGCATACGCATCATCAGCAGCAG
CGAGCAGCAGCAGGCAGCAGCAGCGAGCAGCATACGACTACGAC
TCAGACTACCAACGTTAGCAGCAGCAGGACCAGGCAGGAGCGGC
GAGACGATGCACATCAGCATCAGGCAGCATCGAACGTACGGCAC
GAGGCATATACTACTACTCATTCGCAGAGCGACGAAGCTA

- ACTGCGCAGTACGATCGATCGTACGTACGATCGCATCGATCGAT
CGTAGCATCGTACGTACGATCGATCGATCGATCGATCTAGCTACG
ATCGATCGATCGATCATGGCTGACTGACTGACTCGATCGATCGAT
CGATCGATCGTACGTACGTACGTACGTACGTACGACTGATCGTCG
GCGCGCGATCGTACTACGATCGTACGACGTACTCACGATCGGCAT
CGATCTACGATCGATCGATCGAGCTACGTCGTACGATCGTACGGC
GCGATATACGATCGATACTGACTCGTACGATCGATCGTACGTACG

Similarité de séquences (Level 3)

- ATGCATCGCTAACGCATCAGATGGCTGACTGACTGCATTTTCTCT
ACGATCGCATCGATCTCAGCATCAGCATCATGCAGCATCATCTATA
CGACCAGCATACTCAGCAGCAGCATACGCATCATCAGCAGCAGC
GAGCAGCAGCAGGCAGCAGCAGCGAGCAGCATACGACTACGACT
CAGACTACCAACGTTAGCAGCAGCAGGACCAGGCAGGAGCGGCG
AGACGATGCACATCAGCATCAGGCAGCATCGAACGTACGGCACG
AGGCATATACTACTACTCATTCGCAGAGCGACGAAGCTA

- ACTGCGCAGTACGATCGATCGTACGTACGATCGCATCGATCGAT
CGTAGCATCGTACGTACGATCGATCGATCGATCTAGCTACG
ATCGATCGATCGATCATGGCTGACTGACTGACTCGATCGATCGAT
CGATCGATCGTACGTACGTACGTACGTACGACTGATCGTCG
GCGCGCGATCGTACTACGATCGTACGACGTACTCACTATCGGCAT
CGACCTACGATCGATCGATCGAGCTACGTCGTACGATCGTACGGC
GCGATATACGATCGATACTGACTCGTACGATCGATCGTACGTACG

Similarité de séquences (Niveau 3)

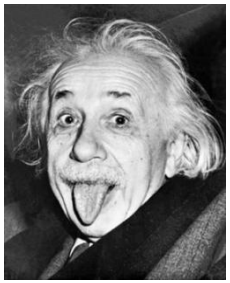
.ATGCATCGCTAACGCATCAGATGGCTGACTGACTGCATTTTCTCT
ACGATCGGCATCGATCTCAGCATCAGCATCATGCAGCGATC
TACGACCAGCATACTC

En pratique:

- Deux genomes complets
- Un génome contre une banque
- Plus de deux génomes

ACTGACTGACTCGATCGATCGAT
CGATCGATCGTACGTACGTACGTACGTACGTACGTACGACTGATCGTCG
GCGCGCGATCGTACTACGATCGTACGACGTACTCACGATCGGCAT
CGATCTACGATCGATCGATCGAGCTACGTCGTACGATCGTACGGC
GCGATATACGATCGATACTGACTCGTACGATCGATCGTACGTACG

Similarité : comparer 2 séquences

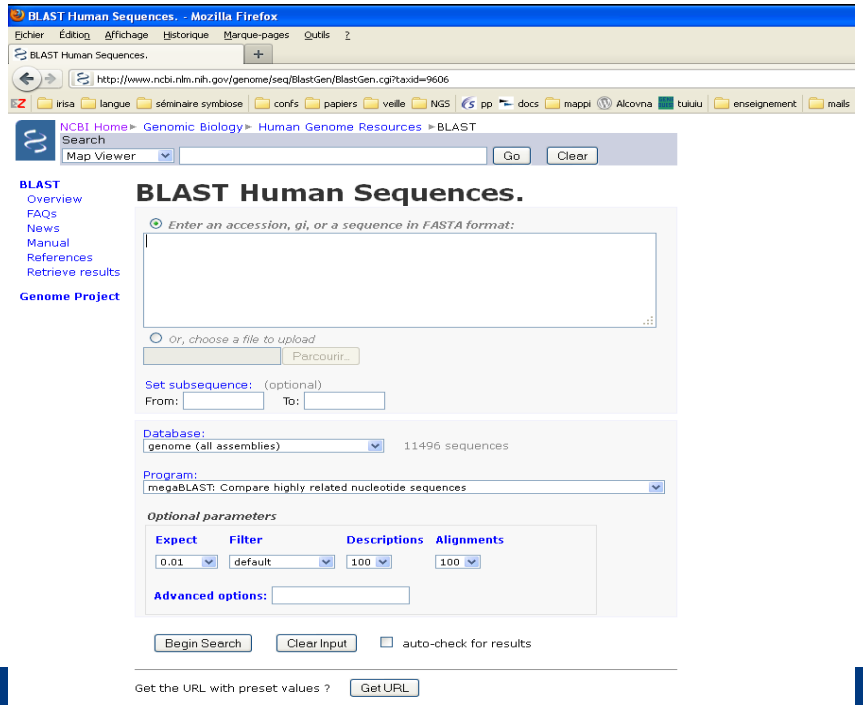


?

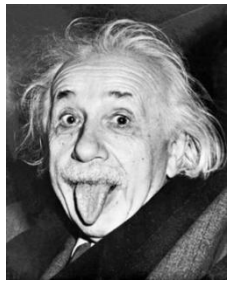
- Trouver les similarités entre deux génomes:
  - 95 ans sur un ordinateur classique actuel
  - Besoins algorithmes
  - Besoins machines spécialisées

Similarité : comparer 2 séquences

- Blast: www.ncbi.nlm.nih.gov/BLAST/
  - Outil de comparaison de séquences.
  - Article scientifique le plus cité, toutes disciplines confondues !

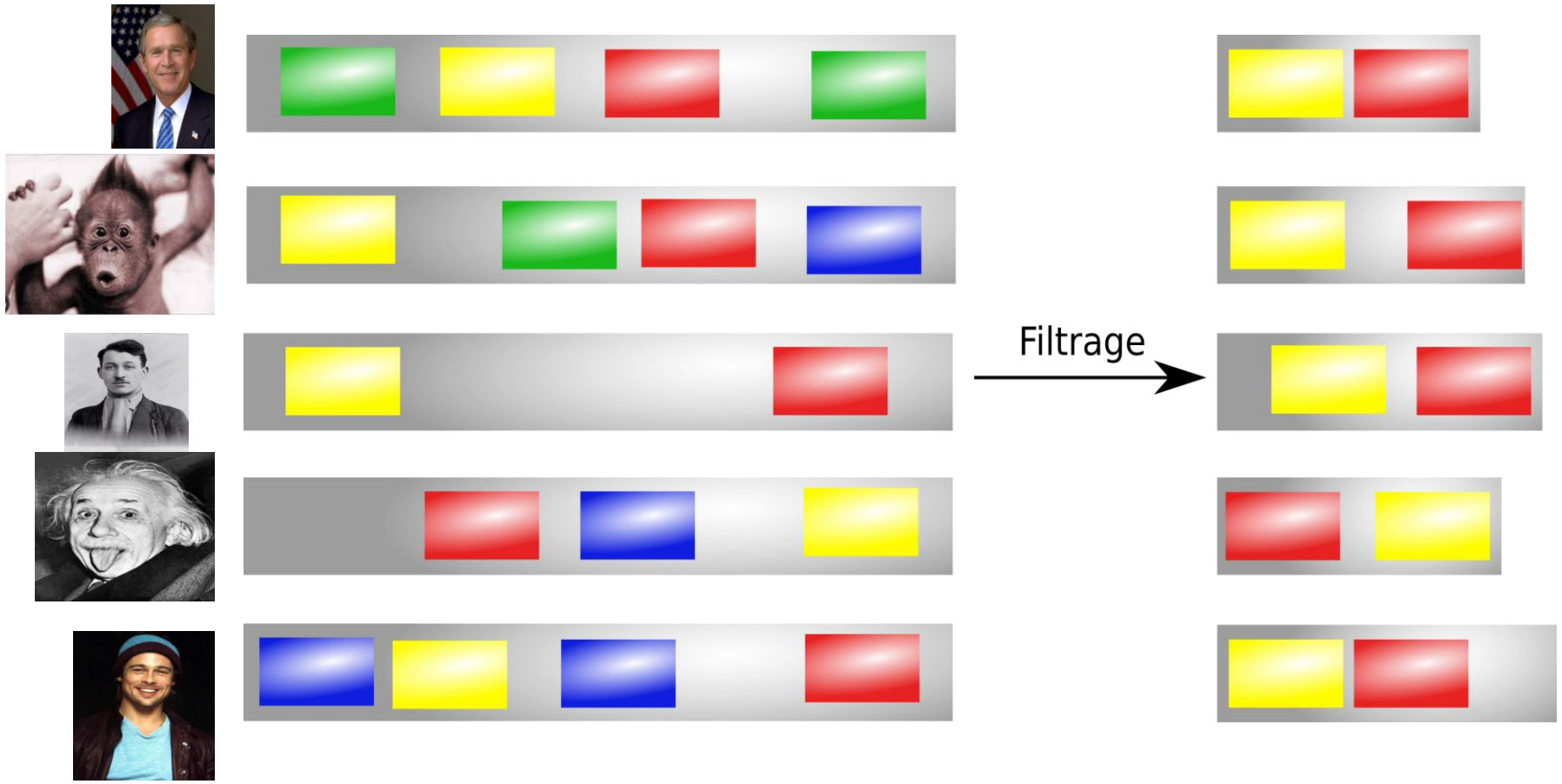


Similarité : comparer plus de 2 séquences



- Similarités de 5 génomes
- 6.10^{19} x âge univers

Similarité : comparer plus de 2 séquences



Les étapes de la construction du filtre.

- Conception modèle (Biologistes)
- Conception méthode (Condition nécessaire mais non suffisante)
- Conception algorithme (Des concepts au langage algorithmique)
- Codage (Ecriture du programme)
- Tests (programme seul / résultats biologiques)
- Diffusion
  - Article
  - Site Web

Similarité : comparer plus de 2 séquences

- Résultats
- Exemple : 24 minutes au lieu de 11h30
- Env. 10 utilisations quotidienne

Lossless filter for multiple repeats with bounded edit distance

Pierre Peterlongo*[1], Gustavo Akio Tominaga Sacomoto[2], Alair Pereira do Lago[3], Nadia Pisanti[4], Marie-France Sagot[5]

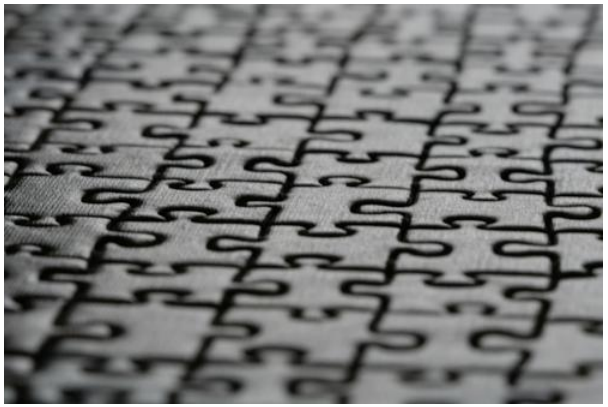
[1]Équipe-projet Symbiose, IRISA / CNRS, Campus de Beaulieu, Rennes, France
[2]Curso Experimental de Ciências Moleculares da Universidade de São Paulo, Brazil
[3]Instituto de Matemática e Estatística da Universidade de São Paulo, Brazil
[4]Dipartimento di Informatica, Università di Pisa, Italy
[5]Équipe BAOBAB, Laboratoire de Biométrie et Biologie Evolutive (UMR 5558); CNRS; Univ. Lyon 1, Villeurbanne Cedex, France *and* Équipe-Projet BAMBOO, INRIA Rhône-Alpes, France *and* King's College, London, UK

Email: Pierre Peterlongo*- pierre.peterlongo@irisa.fr; Gustavo Akio Tominaga Sacomoto - sacomoto@gmail.com; Alair Pereira do Lago - alair@ime.usp.br; Nadia Pisanti - pisanti@di.unipi.it; Marie-France Sagot - Marie-France.Sagot@inria.fr;

*Corresponding author

Autre exemple: Assembler des données

Assembler

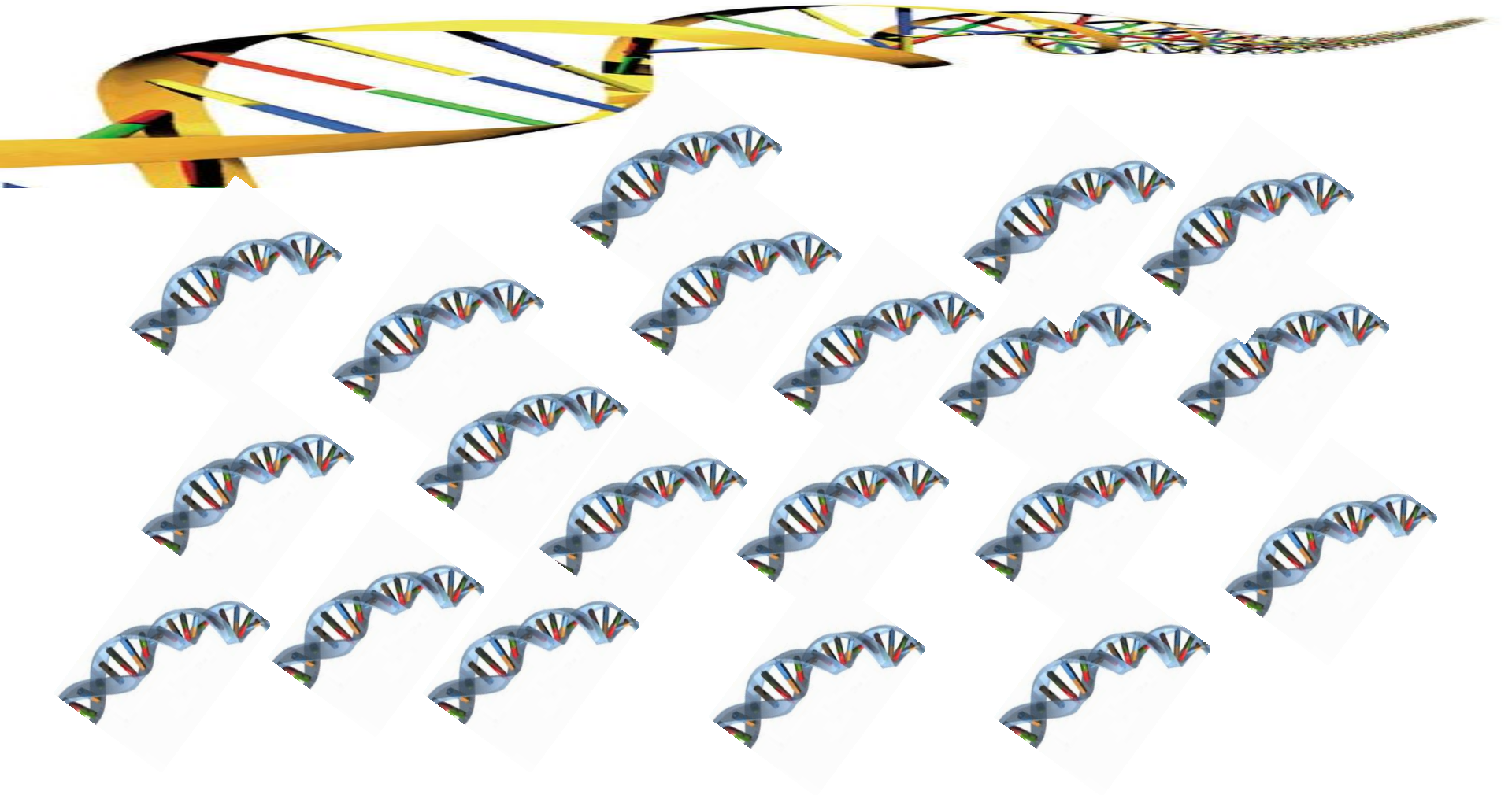


Les défauts des séquenceurs...

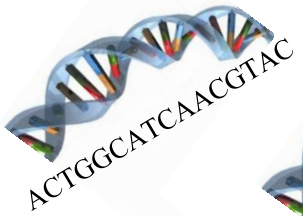
AACTGACGACGC
ATCTACGACTCAGC
AGCAGGCAGCAGCAGC
AGGCAGCAGCAGCAGCATCAGCA
TAACGACTACGCAGGCAGCATCAGCAGACGACTACGA
CTCAGCAGCAGGCACTAACAGCAGCAGCAGCTAGACGCAGCATCATACCGAGCAGCAGCATACGACGCATACGGACCGAGACTA

Les défauts des séquenceurs...

Les défauts des séquenceurs...



Les défauts des séquenceurs...



Que faire de 2 fragments ?

AGTATAGGAAGATAGACAGCAAGC

GAAGATAGACAGCAAGCAGAGATAG

Que faire de 2 fragments ?

AGTATAG**GAAGATAGACAGCAAGC**

GAAGATAGACAGCAAGCAGAGATAG

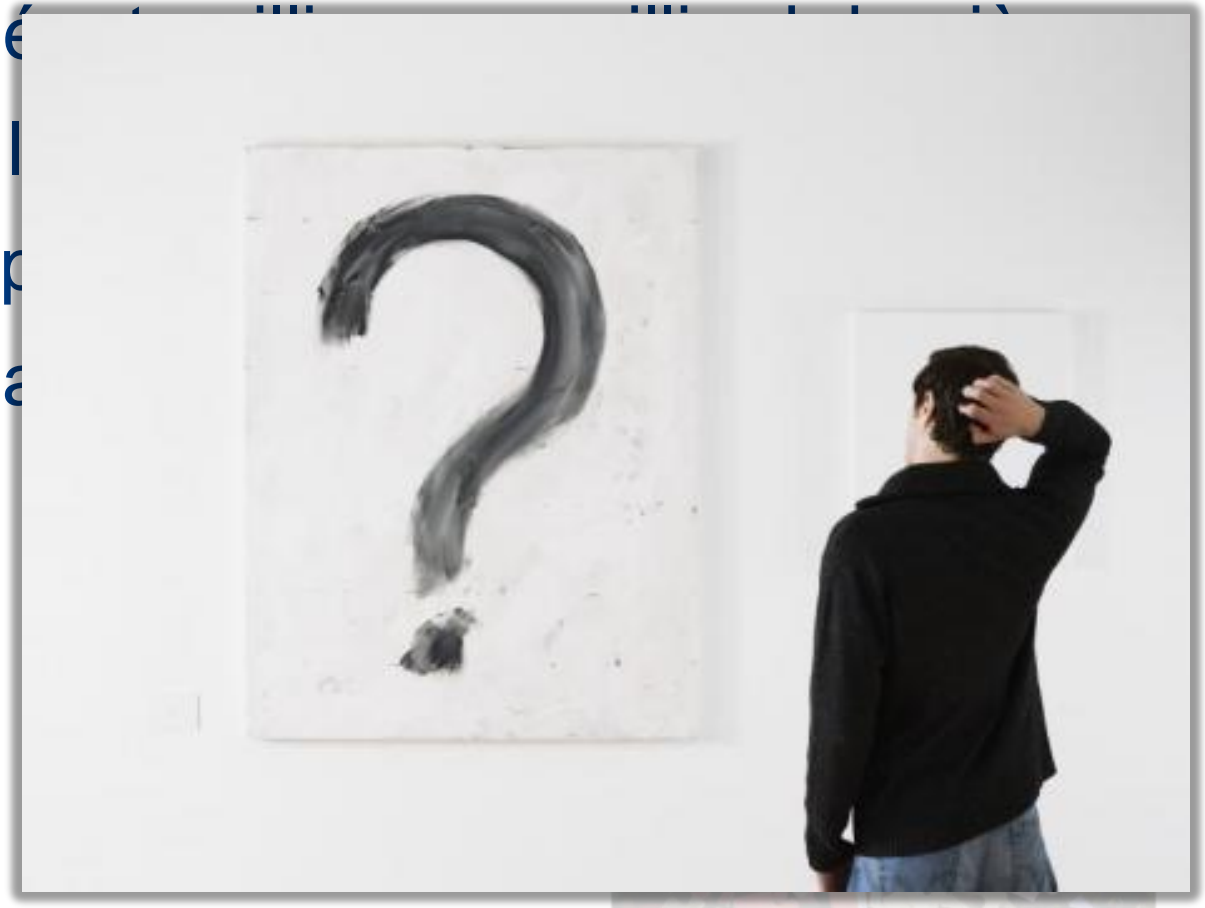
AGTATAG**GAAGATAGACAGCAAGC**AGAGATAG

En pratique: millions/millards de fragments

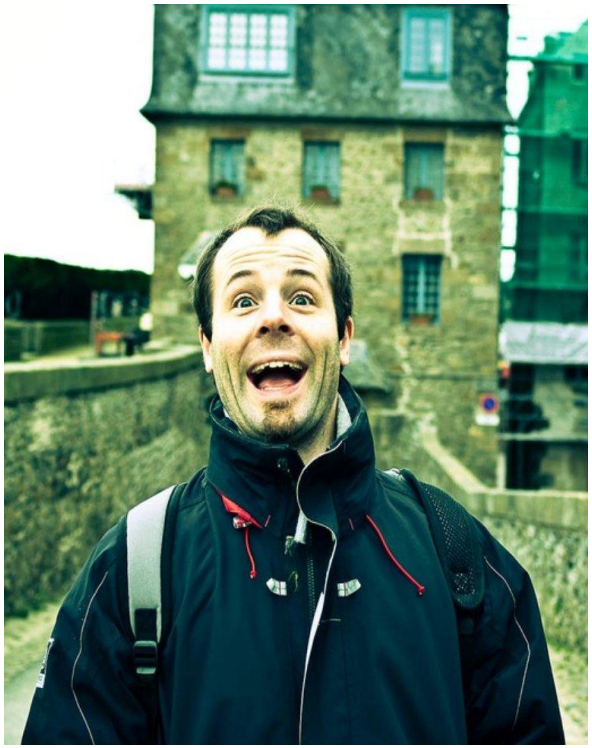
- Puzzle géant: millions ou milliard de pièces
- Manque la boîte avec l'image
- Chaque pièce: pile ou face ?
- Assemblage des pièces imparfait (erreurs)
- ...

En pratique: millions/millards de fragments

- Puzzle gé
- Manque l
- Chaque p
- Assembla
- ...



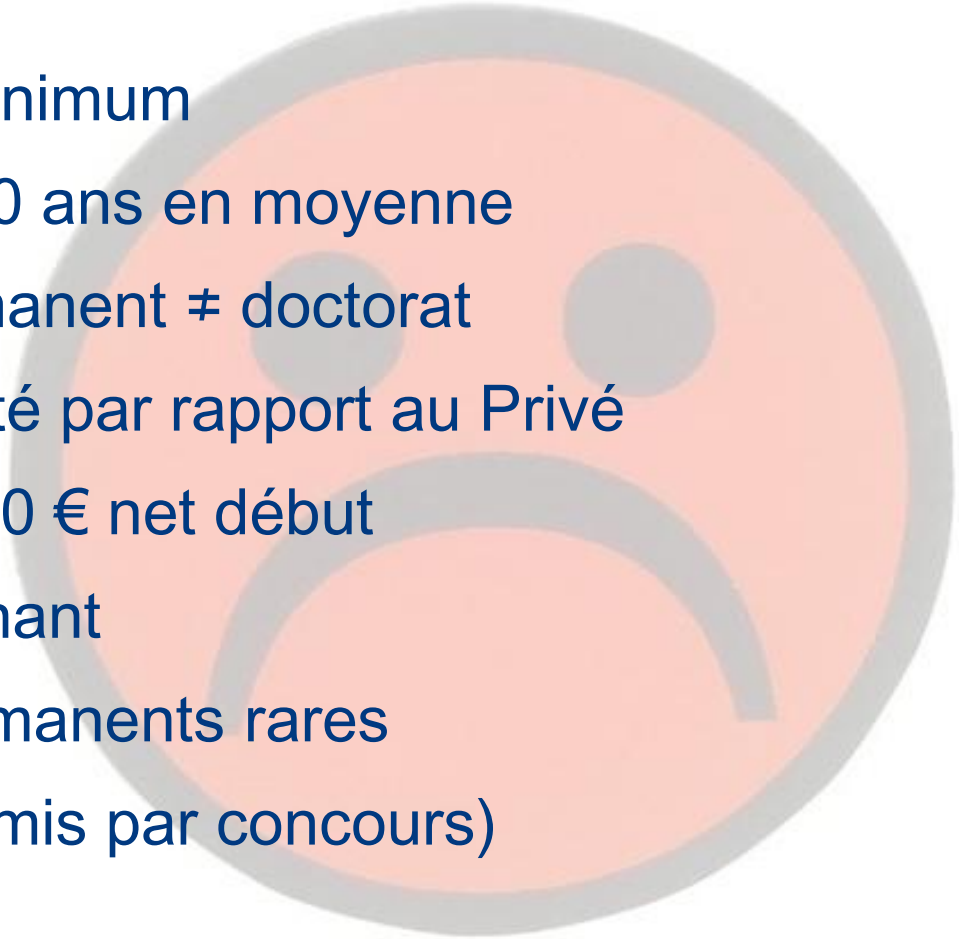
C'est quoi un Bioinformaticien ?



(mon) Parcours

- BAC
- Licence 3 ans (4 en fait)
- Master 2 ans
- Doctorat (thèse) 3 ans 
- Post Doc 2 ans 
- « Chargé de recherche »
  - Concours
  - Fonctionnaire

Les « moins »



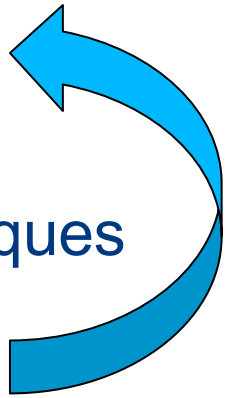
- BAC + 8 minimum
- Fixé vers 30 ans en moyenne
- Poste permanent ≠ doctorat
- Salaire limité par rapport au Privé
- 1600 à 2000 € net début
- Travail prenant
- Postes permanents rares
  - (5 % admis par concours)

Les « plus »

- Plus de « plus » que de « moins »
- Travail passionnant
- Voyages
- Liberté (thématiques)
- Sécurité emploi
- Peu de pression par l'argent
- De moins en moins vrai
- Rencontres (internationales)

Le quotidien

- Derrière un ordinateur (pas de manipulations)
  - Question biologique
  - Inventer de nouveaux algorithmes / logiciels
  - Tester ces logiciels sur des données biologiques
  - Réponse... et d'autres questions



- Discussions avec :
  - des biologistes
  - des informaticiens
  - etc.

Pas le même langage !

Le quotidien, en plus de la « recherche »...

- Recherche dans un contexte international, et collaboratif
  → échanges
  - Publications et exposés dans des conférences
  - Lire / assister / évaluer
  - Écrire / présenter
  - Mais aussi : relire (évaluations), organiser, etc.
  - Bien sûr, en anglais !
- Grand public
- Chercher de l'argent
- Encadrer des étudiants / enseigner à l'université

Merci !

Contact: pierre.peterlongo@inria.fr

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
RENNES - BRETAGNE ATLANTIQUE