

Dans la peau d'un.e Biologiste Computationnel.le ?

Plusieurs casquettes : (Nature Biotechnology, Vol. 31, Number 11, Nov. 2013)

- « - data analyst
- data curator
- database developer
- statistician
- mathematical modeler
- bioinformatician
- software developer
- ontologist
- and many more »

<https://collections.plos.org/collection/translational-bioinformatics/>

Une certitude :

- « *computers are now essential components of modern biological research* »

Une problématique :

- dans quelle limite (éthique par exemple : biohacking in Socialter N°9 Février 2015)

Un besoin :

- un scientifique (vous) est amené à adopter de nouvelles compétences en biologie computationnelle et maîtriser une nouvelle terminologie

De novo transcriptome assembly	is the method of creating a transcriptome without the aid of a reference genome.
The transcriptome	is the set of all RNA molecules, including mRNA, rRNA, tRNA, and other non-coding RNA transcribed in one cell or a population of cells. It differs from the exome in that it includes only those RNA molecules found in a specified cell population, and usually includes the amount or concentration of each RNA molecule in addition to the molecular identities. http://en.wikipedia.org/wiki/Transcriptome
La mégabase (Mb)	est une unité de mesure en biologie moléculaire représentant une longueur de un million de paires de bases d'ADN ou d'ARN.
Le génome	est l'ensemble du matériel génétique d'un individu ou d'une espèce codée dans son acide désoxyribonucléique (ADN) à l'exception de certains virus dont le génome est porté par des molécules d'acide ribonucléique (ARN). Il contient en particulier toutes les séquences codantes (transcrites en ARN messagers, et traduites en protéines) et ARN non codantes (non transcrites, ou transcrites en ARN, mais non traduites). http://fr.wikipedia.org/wiki/G%C3%A9nome

La connaissance (K) biologique est vitale pour l'interprétation des résultats computationnels

- <https://politicalmethodology.wordpress.com/2013/04/01/p-values-are-possibly-biased-estimates-of-the-null-probability/>
- <https://med.stanford.edu/profiles/john-ioannidis?tab=publications>

Soyez un détective des données et de leur manipulation

- <http://www.biomedcentral.com/content/pdf/1471-2105-5-80.pdf>

Utilisez les nombreuses ressources intelligemment et précautionneusement

- http://www.open-bio.org/wiki/Main_Page
- <http://www.biostars.org>
- <http://seqanswers.com>

Comprendre les outils logiciels à disposition :

- ces logiciels sont souvent l'implémentation d'un algorithme générique mais adapté aux données à traiter et qu'il faut être capable de comprendre

Ex. : *Overlap-Layout-Consensus assembler* optimisé pour de longues séquences vs. *Graphes de de Bruijn* pour de courtes séquences

Biologiste ou programmeur :

- ayez une méthode de développement de votre outil (script, pipeline, software) avec des tests sur des cas maîtrisés de données

- utilisez vos connaissances biologiques au mieux notamment pour la documentation

+ des réflexes de bons sens et de génie logiciel à avoir :

- versionnage (Github, Subversion)

- README

- publication conjointe de vos scripts, données et méthodes documentés et commentés (open data, open science, open source, reproductibilité)

- cahier de laboratoire 2.0

Soyez vigilants sur vos découvertes :

- le récent article sur les conclusions faussement statistiquement validées (faux positifs, faux négatifs, valeur P) : prendre en compte le biais, le bruit dans les données introduites au cours des expériences pour les générer ou de leur analyse

Table 1 Essential tools for the biological software developer

Task	Tools
Collaborative software development	Share data and code through online collaborative working environments such as Github, Sourceforge and Bitbucket. Use Google to find tutorials on these systems, e.g., http://try.github.io/
Build powerful pipelines	There are modern software libraries, such as Ruffus, and more traditional tools, such as Make, to build pipelines from existing software tools. Your choice will depend on personal preference and on your favorite programming language.
Make your pipelines available	You may be comfortable on the command line, but your collaborators may not be. Therefore you can deliver your pipelines through graphical environments such as Galaxy (http://www.galaxyproject.org/) or Taverna (http://www.taverna.org.uk/).
Integrated development environment (IDE)	Whether you want to adopt a full IDE, such as Eclipse, or an advanced text editor, such as Emacs, you will need something to use to develop your code. Again, this will likely depend on your choice of language and personal preference. However, at some point, you'll have to use a command line-based editor, such as vim or nano, so it's advisable to learn at least the basics.

Table 2 Useful resources for learning

Type of information	Relevant URLs
MOOCs (massive open online courses)	These are very popular at the moment and offer free training over the internet. Coursera (https://www.coursera.org/), Udacity (https://www.udacity.com/), edX (https://www.edx.org/) and the Kahn Academy (https://www.khanacademy.org/) have a range of courses relevant to bioinformatics, genomics, computing, statistics and modeling.
Learning to code	Codecademy (http://www.codecademy.com/) and Code School (https://www.codeschool.com/) are not specific to biology but do offer simple ways to learn how to code. For a more biological perspective, “Python for biologists” (http://pythonforbiologists.com/) is always popular. For examples of best practices visit http://software-carpentry.org/ .
Bioinformatics problem solving	Learn bioinformatics through problem solving and pit your wits against others at http://www.rosalind.info .
Web forums	These are essential when you start out—ask questions and receive answers from experts at http://www.seqanswers.com/ and http://www.biostars.org/ .
International organizations	GOBLET is the global organization for bioinformatics learning education and training (http://www.mygoblet.org/), and ELIXIR is a European organization set up to provide an infrastructure, including training, for life sciences information (http://www.elixir-europe.org/).
Blogs and lists	A variety of blogs and lists exist online that detail computational biology courses, such as http://stephenturner.us/p/edu and http://ged.msu.edu/angus/bioinformatics-courses.html .

Les Biotechs : nouvelle révolution industrielle ? Eldorado pour les laboratoires ?

Démocratisation du séquençage ADN depuis 2012 :

- 1^{er} génome humain décodé : 15 ans de recherche, 2.7 milliards de dollars
- Gain de 10^5 pour le séquençage de mégabase et de 10^4 pour un génome (quelques milliers de dollars et d'heures contre 100 millions en 2001)

Post-génomique :

- biohacking (DIY bio), virus intelligents (*Microbesoft* ?), biologie de synthèse (OGM 2.0?) (marché de 1 000 milliards de dollars en 2025, Source OCDE)
- réflexion éthique : ONG ETC group (<http://www.etcgroup.org/fr>), Fondation Sciences Citoyennes



ANTIDOTE

CONTRE LE VIEILLISSEMENT

Google a fondé **Calico**, entreprise de biotech spécialisée dans les recherches sur le vieillissement et regroupant des experts en génie génétique. La société tente de fabriquer des médicaments contre les maladies de Parkinson et d'Alzheimer. Objectif ultime : trouver un remède contre le vieillissement.

DIAGNOSTIC PRÉDICTIF

Le projet **Baseline Study** vise à dresser la carte génétique type d'un individu en bonne santé, en collectant les génomes de milliers de volontaires. En analysant ces données, Google espère pouvoir détecter à terme les risques de maladies graves, pour les soigner avant qu'elles ne se déclarent.



DES ANTICORPS THÉRAPEUTIQUES

Adimab conçoit des immunoglobulines. Ces molécules sont conçues par les globules blancs, mais la société les fabrique à partir de souches de levures et d'ADN. Ses spécialistes en biologie structurale créent des « systèmes immunitaires synthétiques » - des « anticorps thérapeutiques » qui stimulent les antigènes liés à des maladies infectieuses ou au cancer. Financement : 14 millions de \$



GÉNOMIQUE PERSONNELLE

Google a investi 4 millions de \$ dans **23andMe**, spécialisée dans l'analyse du code génétique. La start-up propose des tests permettant de déceler des risques de maladies dans des séquences d'ADN. Son outil de « diagnostic de maladie » (interdit aux États-Unis mais commercialisé au Royaume-Uni) permet de détecter des facteurs de risque liés à la maladie de Parkinson ou au cancer du sein.

La galaxie génomique de



LE GÉNIE GÉNÉTIQUE CONTRE ALZHEIMER

Google soutient, à hauteur de 22 millions de \$, l'entreprise biopharmaceutique **Iperian**, spécialisée dans la recherche sur les maladies dégénératives. Son but est de concevoir, à partir de cellules souches pluripotentes (fabriquées et reprogrammées en laboratoire), des médicaments s'attaquant aux maladies en les modifiant. Iperian espère ainsi créer un système de « thérapie cellulaire » face à la maladie d'Alzheimer.

FLATIRON

LE « LEXIS-NEXIS DU CANCER »

Flatiron Health collecte des informations sur les personnes atteintes de cancer, afin de développer des « pipelines de données » destinés aux chercheurs et aux médecins. C'est le plus gros investissement de Google en santé, avec 130 millions de \$. La plateforme de Flatiron, « Oncology Cloud », utilisée par 2 000 cliniciens, comprend une base de données de milliers de dossiers médicaux électroniques.

DNAexus

LE BIG DATA ADAPTÉ AU SÉQUENÇAGE

DNAexus, en Californie, est spécialisée dans le big data appliqué au séquençage de l'ADN. Elle fournit des services permettant de traiter des données de séquençage, stockées sur internet. L'objectif de DNAexus est de sauvegarder votre séquençage ADN sur le Cloud, puis de vous proposer de l'analyser, en dressant votre profil génétique. Google finance DNAexus à hauteur de 15 millions de \$.



PILULE INTELLIGENTE

Rani Therapeutics développe une pilule capable de délivrer de l'insuline dans l'organisme, remplaçant ainsi les piqûres auxquelles sont habitués les diabétiques. Cette pilule « intelligente » pourrait aussi s'appliquer, promet Rani Therapeutics, à la sclérose en plaques. Financement : 10 millions de \$.

Start-up :

Transgène : virus intelligents contre cancer poumon et foie

Collectis : cellules immunitaires génétiquement modifiées (France avec Pfizer)

Amyris : fabrication artificielle de l'artémisinine (anti-paludique) (US avec Sanofi)

Global Bioenergies : biocarburant (Evry avec Audi)

Abolis Biotechnologies : CAO de microorganismes

Hyasynth Bio : optimiser la production de cannabinoïdes à visée médicale (Canada)

Biotechnologies :	Utilisation de techniques utilisant des être vivants, en général modifiés génétiquement, pour la fabrication industrielle de médicaments, de matières premières industrielles, ou pour améliorer la production agricole
Génie génétique :	Techniques reposant sur la génétique pour utiliser, reproduire ou modifier le génome des êtres vivants
Transgénèse :	Introduction d'un ou plusieurs gènes dans un organisme vivant, afin de le modifier (OGM)
Séquençage de l'ADN :	Détermination de l'ordre d'enchaînement des nucléotides constituant un fragment d'ADN
Biologie de synthèse :	Conception (ingénierie) de composants et de systèmes biologiques qui n'existent pas dans la nature (artificiels), ou modifications d'éléments biologiques existants
Biohacking :	Pratique visant à combiner biotechnologies (expérimentation génétique de l'ADN et du vivant) et « hacking » (bricolage) pour expérimenter les possibilités du génome En contrepoint des ONG : Fondation Sciences Citoyennes, ETC group (http://www.etcgroup.org/fr)

Un spectre de disciplines si large

Computational biophysics & Structural Biology

Modeling of Regulatory, signaling and metabolic networks

Pattern Recognition and Machine Learning

Data Mining / Graph Mining / Sequence Mining

Functional Genomics

Molecular Interaction **Networks** / Systems Biology + Structural Biology

Prediction of Protein-Protein and Protein-DNA interactions

Gene Expression Analysis & Prediction of Regulatory Network Structure

Study of Complex Inherited Traits

Image Analysis & Interpretation

Biomedical Ontology Development

Knowledge Extraction From Scientific Litterature & Medical Reports

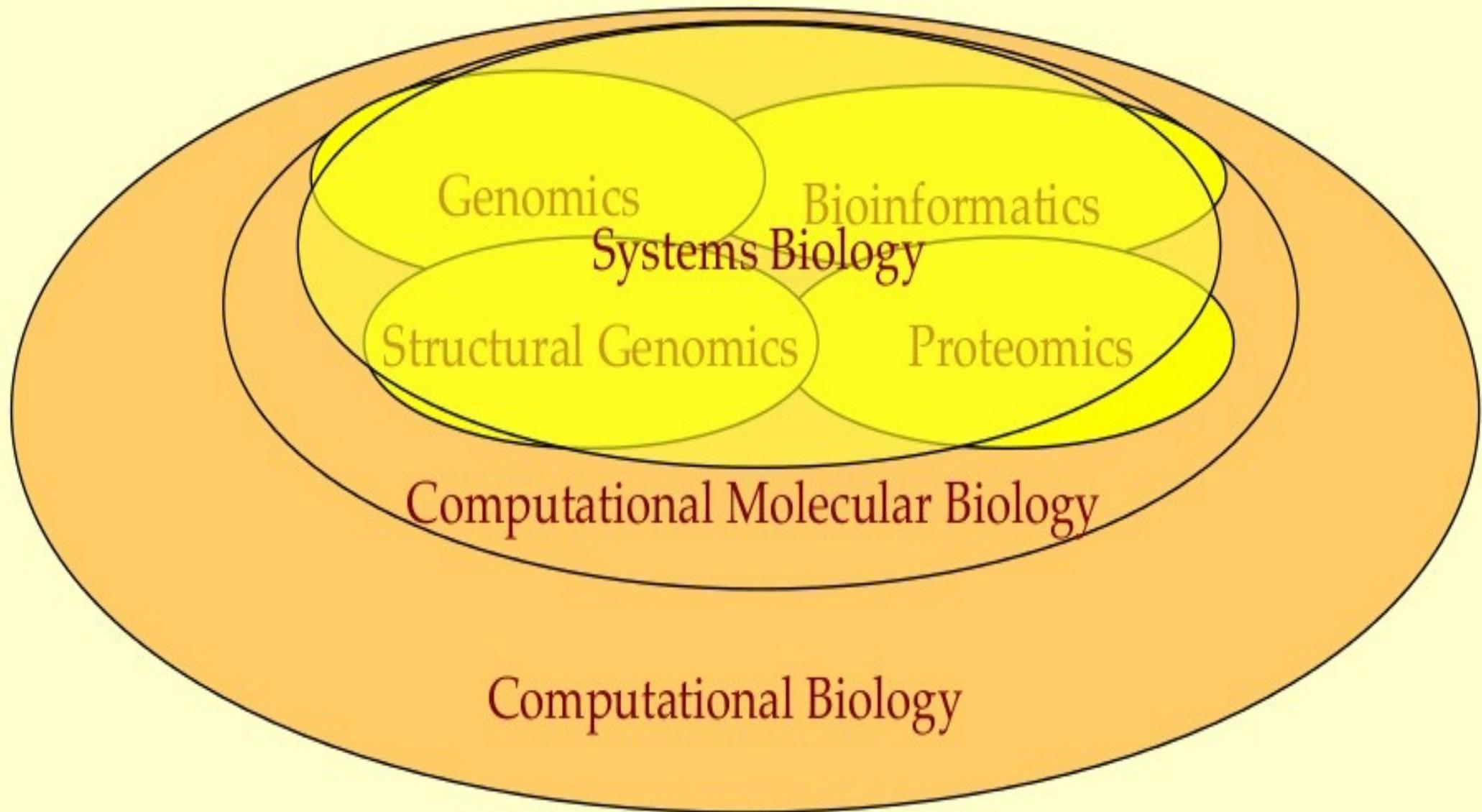
Evidence Integration

Protein Structure Modeling

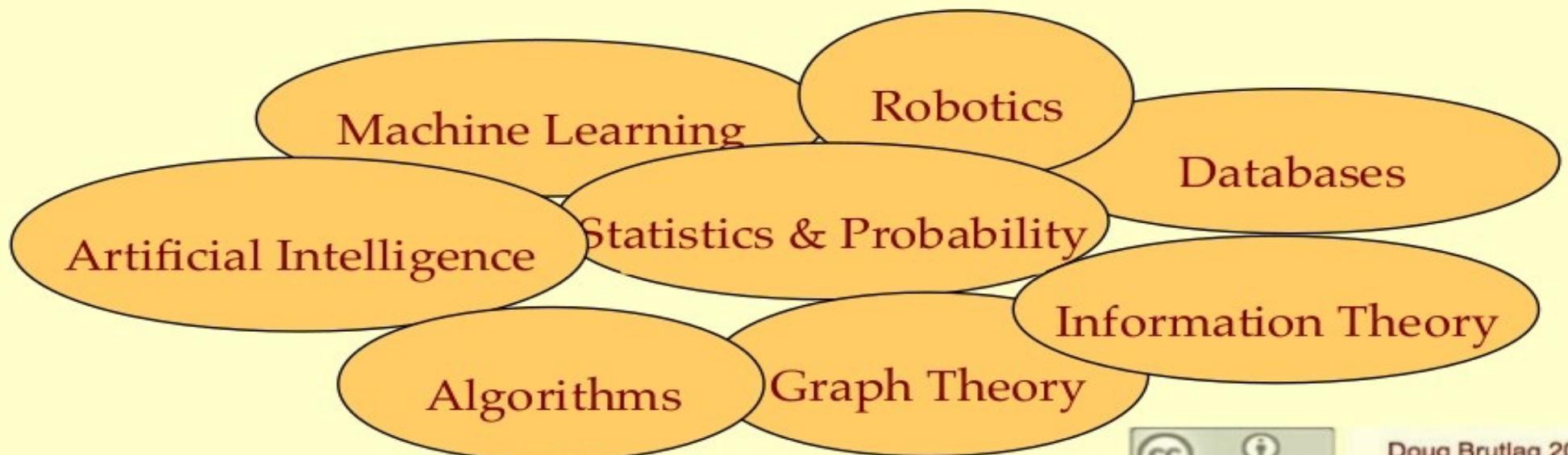
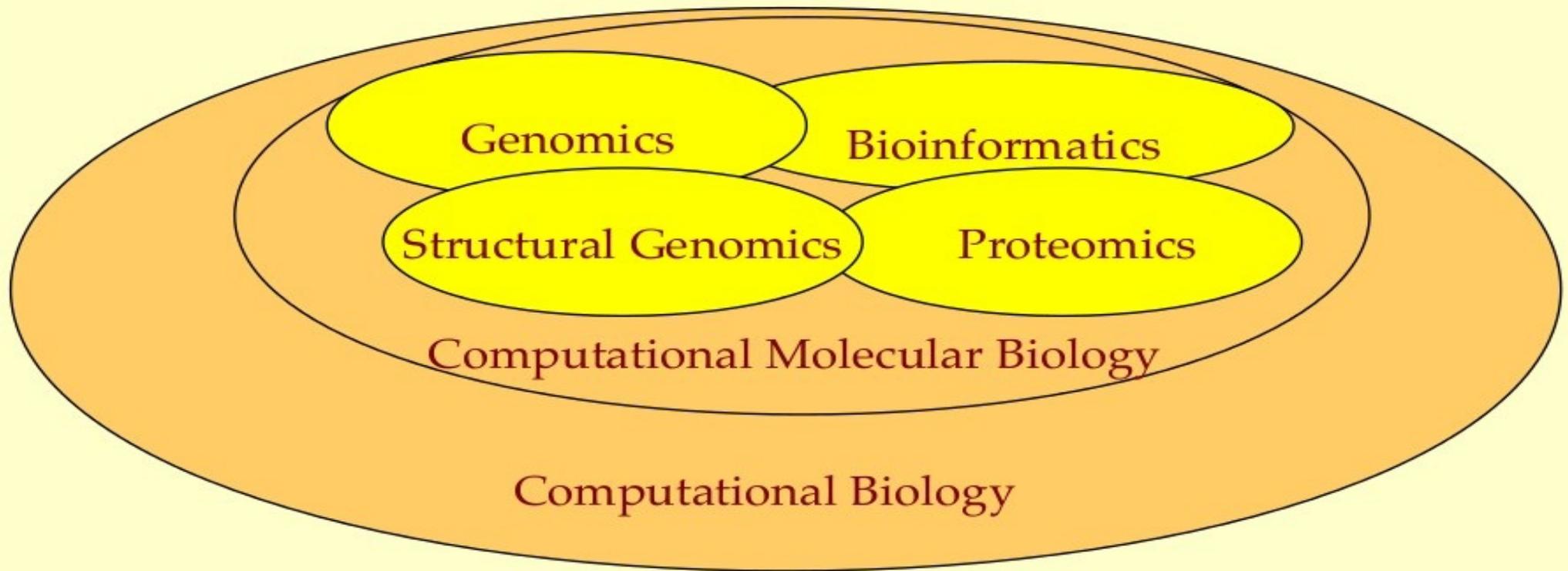
Phylogenetic Tree Construction

Synthetic Biology

Genomics, Bioinformatics & Computational Biology



Genomics, Bioinformatics & Computational Biology



Biologie et *computer sciences* : l'interdisciplinarité à l'état brut

Variabilité génétique chez l'homme : 0.1 % (vs. 0.2 % chez le chimpanzé)

*Longueur moyenne d'une chaîne protéique : 300-500 acides aminés.
La plus longue : la protéine titine humaine avec 34 350 acides aminés,
impliquée dans l'élasticité de la fibre musculaire*



pangolin



tatou



félin

Biologie et *computer sciences* : l'interdisciplinarité à l'état brut

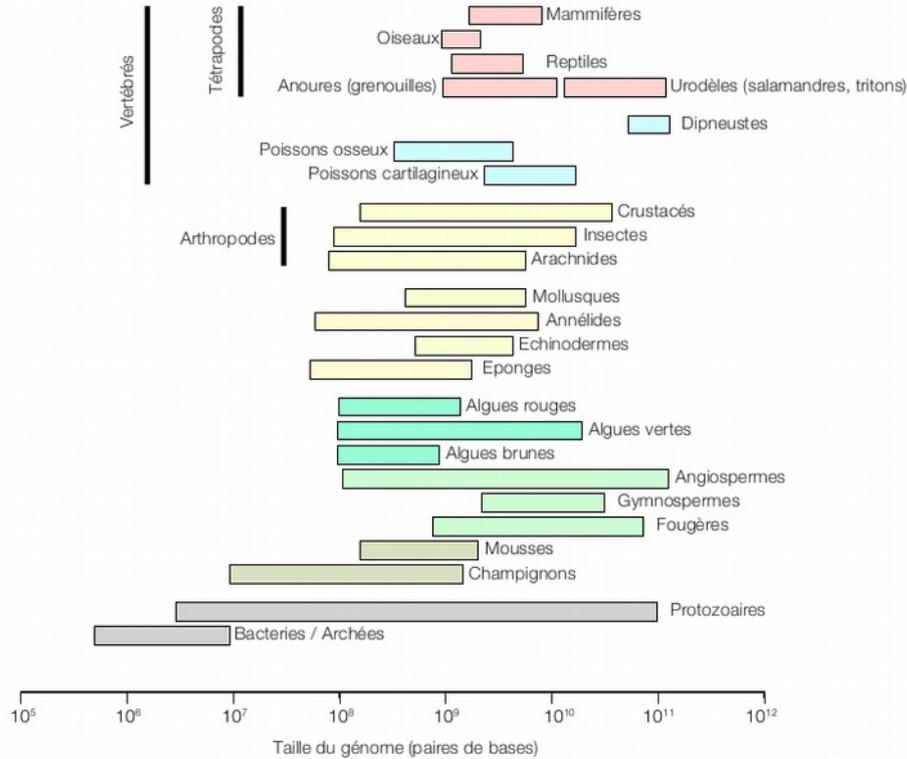


Figure 5.1 : Variabilité de la taille des génomes chez différentes familles d'organismes vivants. Pour chaque groupe indiqué, la boîte indique l'étendue de la fourchette des tailles de génome observées. Notez que l'échelle est logarithmique. Pour les angiospermes (plantes à fleurs), par exemple, la variation est donc dans un rapport de 1 à 1000 (10^9 à 10^{11} pb). Données www.genomesize.com (animaux), <http://data.kew.org/cval>

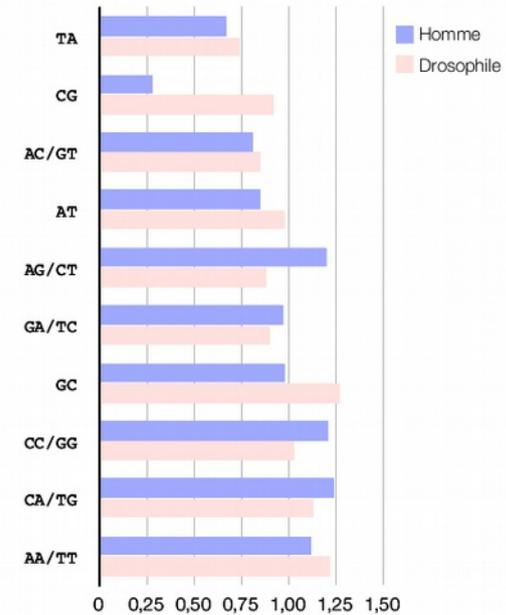


Figure 7.6 : Fréquences normalisées des différents dinucléotides dans les génomes de l'Homme et de la drosophile. Ces fréquences étant calculées sur les deux brins d'ADN du génome, on a affiché de manière regroupée les dinucléotides complémentaires. Ainsi, par exemple, la fréquence de AC et celle de GT sont identiques, puisqu'il s'agit de séquences complémentaires. Chaque fois que l'une apparaît sur un brin, la seconde apparaît sur l'autre, si bien que leur nombre d'occurrences, et donc leur fréquence, sont identiques. Comme il s'agit de fréquences normalisées, la valeur attendue pour une séquence aléatoire, sans biais, est 1,00. Les dinucléotides du bas du graphe sont sur-représentés ($> 1,00$), ceux du haut sont sous-représentés ($< 1,00$).

Biologie et *computer sciences* : l'interdisciplinarité à l'état brut

Organisme	Taille du génome (pb)
Virus du SIDA	9 750
<i>Mycoplasma genitalium</i>	580 000
<i>Helicobacter pylori</i> (ulcère stomacal)	1 667 867
<i>Escherichia coli</i>	4 639 221
Levure de bière	12 067 280
<i>Plasmodium falciparum</i> (paludisme)	25 000 000
Trypanosome	35 000 000
Nématode	110 000 000
Drosophile	150 000 000
Tétraodon (poisson-zèbre)	350 000 000
Tomate	655 000 000
Soja	1 115 000 000
Poulet	1 200 000 000
Boa constrictor	2 100 000 000
Homme	3 400 000 000

Organisme	Nombre de gènes	Taille du génome (Mb)	Densité (gènes/Mb)
<i>Haemophilus influenzae</i> (bactérie)	1 800	1,8	~1 000
<i>Escherichia coli</i> (bactérie)	4 300	4,6	~930
Levure de bière (champignon)	6 000	12,1	~500
Drosophile (insecte)	~14 500	150,0	~100
Nématode (ver)	~21 000	110,0	~190
Arabette (plante)	~25 500	110,0	~230
Souris (mammifère)	~25 000	2 700,0	~9
Homme (mammifère)	~25 000	3 400,0	~7
Paramécie (protiste cilié)	~40 000	72,0	~550

	Bactérie	Drosophile	Homme
Longueur L du génome	10^6 - 10^7	10^8	$3 \cdot 10^9$
Nombre n de fragments séquencés (longueur $k \approx 600$)	10 000-100 000	10^6	$3 \cdot 10^7$
Longueur totale séquencée ($k \cdot n$)	$6 \cdot 10^6$ - $6 \cdot 10^7$	$6 \cdot 10^8$	$1,8 \cdot 10^{10}$
Nombre de comparaisons de fragments ($\sim n^2$)	10^8 - 10^{10}	10^{12}	10^{15}



Prof. Frédéric Dardel

Biologie Cellulaire et Moléculaire /
Modélisation Mathématique



Nicolas Loménie
Université de Paris,
Systèmes Intelligents de Perception
<http://w3.mi.parisdescartes.fr/sip-lab/>

Nicolas.lomenie@u-paris.fr

Emmanuel Logak

Ex-étudiant Info et actuellement Filière médecine