www.facebook.com/DomaineSNV

BI4U2 – Bioinformatique appliquée

http://pedagogix-tagc.univ-mrs.fr/courses/bioinfo_intro/

Domaine SNV : Biologie,Agronomie,Science Alimentaire,Ecologie

Introduction

Jacques van Helden

Jacques.van-Helden@univ-amu.fr
Aix-Marseille Université, France
Technological Advances for Genomics and Clinics
(TAGC, INSERM Unit U1090)
http://jacques.van-helden.perso.luminy.univ-amu.fr/

Matériel de cours

- Diapos, énoncés des TP
  - http://pedagogix-tagc.univ-mrs.fr/courses/bioinfo_intro/
- Dépôt des rapports de TP
  - http://ametice.univ-amu.fr/course/view.php?id=11052

Objectifs pédagogiques

- Ce cours est destiné à des étudiants en sciences de la vie (biologie, biochimie, sciences biomédicales).
- Théorie (10h CM)
  - Introduction aux concepts et méthodes de base en bioinformatique.
  - Illustrations sur base d'exemple concrets.
- Pratique (5x4h TP)
  - Utilisation des outils bioinformatiques pour analyser des séquences biologiques.
  - Interprétation biologique des résultats
  - Evaluation de la fiabilité statistique des alignements de séquences

Qu'est-ce que la bioinformatique ?

Quelques définitions de la bioinformatique (1)

- Les bioinformaticiens définissent leur propre domaine de diverses manières
- Georgia Inst of Tech., USA
  - *“Bioinformatics is an integration of mathematical, statistical and computer methods to analyse biological, biochemical and biophysical data”*
  - *"Intégration des méthodes mathématiques, statistiques et informatiques pour analyser les données biologiques, biochimiques et biophysiques"*
- Cette définition me semble assez pertinente, mais présente la faiblesse d'être motivée par les données plutôt que par les questions.

Quelques définitions de la bioinformatique (2)

- Les bioinformaticiens définissent leur propre domaine de diverses manières
- Stanford University, USA
  - *“Bioinformatics is the study of biological information as is passes from its storage site in the genome to the various gene products in the cell. …it involves the creating and development of advanced information and computational technologies for problems in molecular biology…”*
  - *"La bioinformatique est l'étude de l'information biologique quand elle passe de son site de stockage dans le génome aux différents produits des gènes dans la cellule. […] Elle inclut la création et le développement de technologies informatiques avancées pour les problèmes de la biologie moléculaire.*
- Cette définition me semble trop restrictive. En particulier, "les produits des gènes" réduit le domaine à l'analyse des protéines. La bioinformatique inclut d'autres champs d'application, comme l'étude du métabolisme, des séquences nucléiques, de l'évolution, etc.
  1. *“Bioinformatics specifically refers to the search and use of patterns and structure in biological data and the development of new methods for database access.”*
     - (Virginia Inst Tech., USA)
     - No doubt that this definition was written by a computer scientist, or an informatician, but not by a bioinformatician.

Quelques définitions de la bioinformatique (3)

- Les bioinformaticiens définissent leur propre domaine de diverses manières
- Virginia Inst Tech., USA
  - *“Bioinformatics specifically refers to the search and use of patterns and structure in biological data and the development of new methods for database access.”*
  - *"La bioinformatique se réfère spécifiquement à la recherche et à l'utilisation de patterns et de structures dans les données biologiques et au développement de nouvelles méthodes pour accéder aux bases de données."*
- Sans aucun doute, cette définition a été écrite par un informaticien, et non par un biologiste ou un bioinformaticien.

Quelques définitions de la bioinformatique (4)

- Certains établissent une distinction entre "bioinformatique" et "biologie computationnelle".
- Pour autant que je sache, les deux termes étaient initialement utilisés indistinctement pour désigner la même discipline. Les tentatives ultérieures de délimiter une frontière entre "bioinformatique" et "biologie computationnelle" me semblent quelque peu arbitraires, et vaines.

- Virginia Inst Tech., USA
  - *"Bioinformatics specifically refers to the search and use of patterns and structure in biological data and the development of new methods for database access. Computational biology is more frequently used to refer to physical and mathematical simulation of biological processes."*
  - *"La bioinformatique se réfère spécifiquement à la recherche et à l'utilisation de patterns et de structures dans les données biologiques et au développement de nouvelles méthodes pour accéder aux bases de données. La biologie computationnelle est plus fréquemment utilisée pour se référer aux simulations physiques et mathématiques des processus biologiques."*

- Source: Biomedical Information Science and Technology Initiative http://www.bisti.nih.gov/

Quelques définitions de la bioinformatique (5)

- Certains établissent une distinction entre "bioinformatique" et "biologie computationnelle".
- Pour autant que je sache, les deux termes étaient initialement utilisés indistinctement pour désigner la même discipline. Les tentatives ultérieures de délimiter une frontière entre "bioinformatique" et "biologie computationnelle" me semblent quelque peu arbitraires, et vaines.

- National Institute of Health (NIH), USA. Working Definition of Bioinformatics and Computational Biology - July 17, 2000
  - "Bioinformatics : Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data."
  - *"Bioinformatique: recherche, développement ou application d'outils informatiques [computationnels ?] et d'approches pour étendre l'utilisation des données biologique, médicales, comportementales ou sanitaires, y compris [les outils et approches] pour acquérir, entreposer, organiser, archiver, analyser ou visualiser de telles données."*
  - "Computational Biology: The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems."
  - *"Biologie computationnelle: développement et application de méthodes analytiques et théoriques, de modélisation mathématique et de techniques de simulation informatique [computationnelle ?] pour l'étude de systèmes biologiques, comportementaux et sociaux."*

- Source: Biomedical Information Science and Technology Initiative http://www.bisti.nih.gov/

How would I define it ?

- *Développement et applications de méthodes informatiques, statistiques, mathématiques et physiques pour l'analyse de données biomoléculaires.*
- Development and applications of methods from computer sciences, statistics, mathematics and physics to analyse biomolecular data.

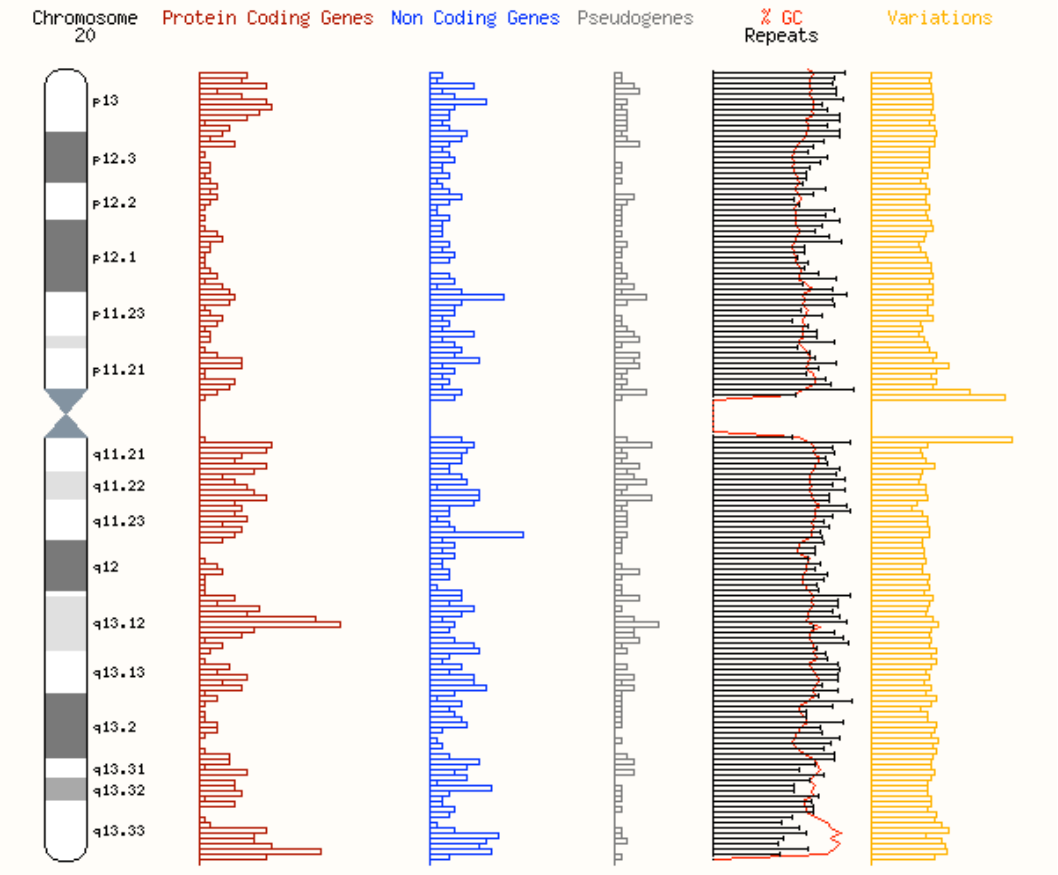
La bioinformatique – pour quoi faire ?

Domaines de la bioinformatique (liste non exhaustive)

- Gestion des données
- Structures moléculaires
  - Visualisation, analyse, classification, prédiction
- Analyse de séquences
  - Alignements, recherches de similarités, détection de motifs
- Génomique
  - Annotation des génomes, génomique comparative
- Phylogénie
  - Relations évolutives entre gènes, entre génomes, entre organismes
  - Inférence de scénarios évolutifs
- Génomique fonctionnelle
  - Transcriptome, protéome, interactome
- Analyse des réseaux biomoléculaires
  - Réseaux métaboliques, d'interactions protéiques, de régulation génétique, …
- Biologie des systèmes
  - Modélisation et simulation des propriétés dynamiques des systèmes biologiques
- …

Analyse des génomes

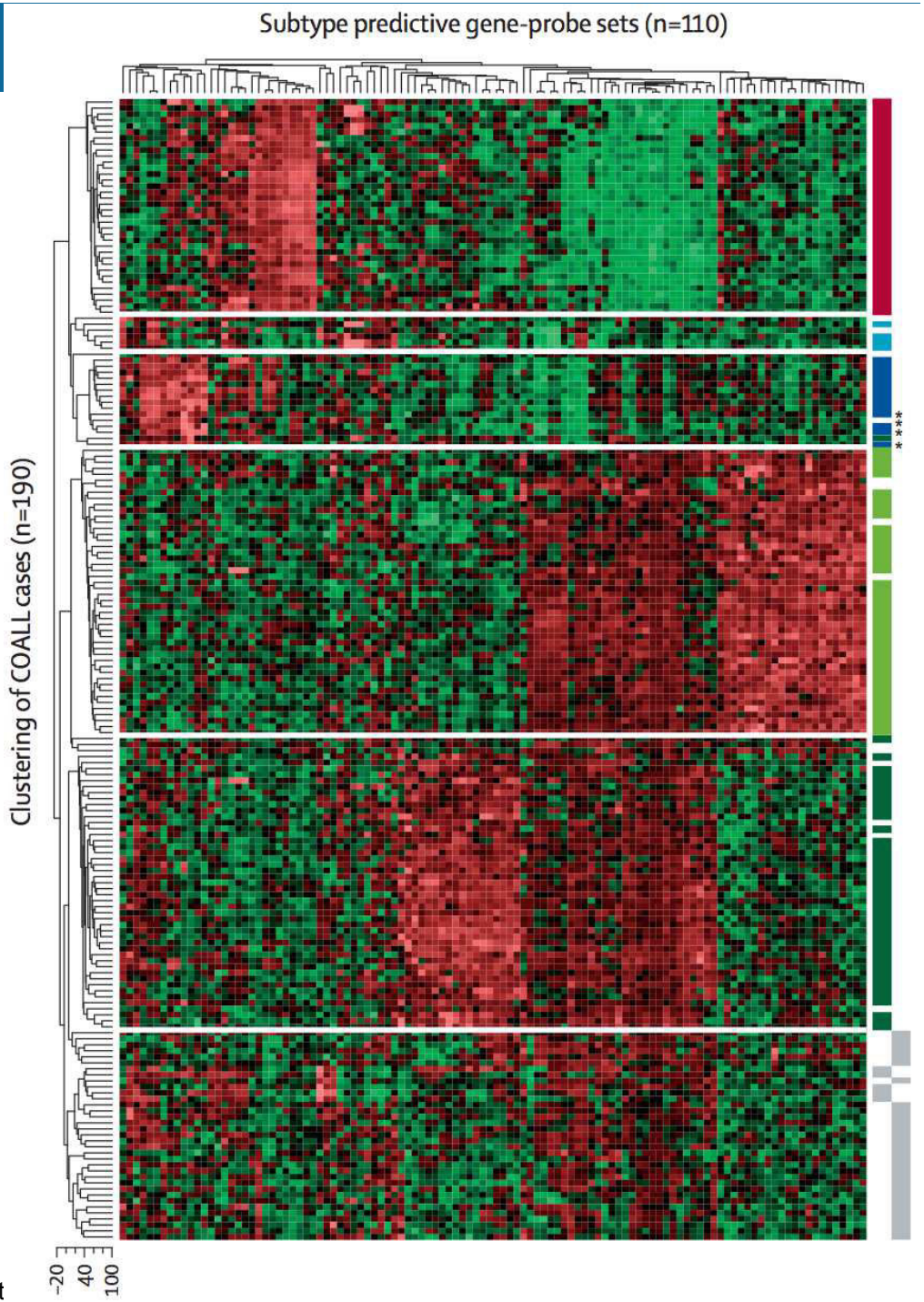
- Exemple: vue schématique du chromosome humain numéro 22 (source: Ensembl).
- La bioinformatique est utilisée à chaque étape d'un projet de séquençage génomique.
  - Stockage des séquences primaires
  - Assemblage des séquences chromosomiques
  - Prédiction de la localisation des gènes
  - Annotation des gènes (prédiction de leur fonction sur base de leur séquence, recherches bibliographiques).
  - Analyse de la composition chromosomique (contenu en GC, variations interindividuelles, …).
  - …



- http://www.ensembl.org/Homo_sapiens/Location/Chromosome?r=20:31817450-31917450

Analyse du transcriptome

- La transcription des gènes est précisément régulée: chaque gène est exprimé à un niveau spécifique en fonction du type cellulaire, du tissu, du temps, des conditions intra- et extra-cellulaires, …
- Depuis 1997, les technologies des biopuces ont été développées pour mesurer les concentrations de tous les ARNs d'une cellule.
- Le transcriptome est défini comme l'ensemble de toutes les molécules d'ARN transcrites à partir d'un génome.
- Depuis 1997, l'analyse du transcriptome a été utilisée pour comprendre les mécanismes de régulation transcriptionnelle, ainsi que pour certaines applications médicales (exemple ci-contre: classification des cancers).
- Figure: classification de leucémies lymphoblastiques aigues en sous-types (lignes) sur base de profils d'expression pour une série de gènes marqueurs (colonnes).



- Figure source: Den Boer et al. A subtype of childhood acute lymphoblast

Le séquençage à très haut débit ("next generation sequencing (NGS)"

- Le coût du séquençage a baissé de façon exponentielle depuis les années 1990, grâce à l'amélioration et à l'automatisation des techniques, stimulées par les projets de séquençage de génomes.
- Jusqu'en 2006, cette décroissance était plus ou moins proportionnelle à la décroissance exponentielle des coûts de stockage et d'analyse informatique (loi de Moore).
- Depuis 2007, plusieurs compagnies ont proposé des nouvelles technologies beaucoup plus rapides. Le coût du séquençage decroît beaucoup plus vite que celui du stockage.
- Les biologistes sont confrontés à un réel problème pour stocker et analyser les données qu'ils produisent.

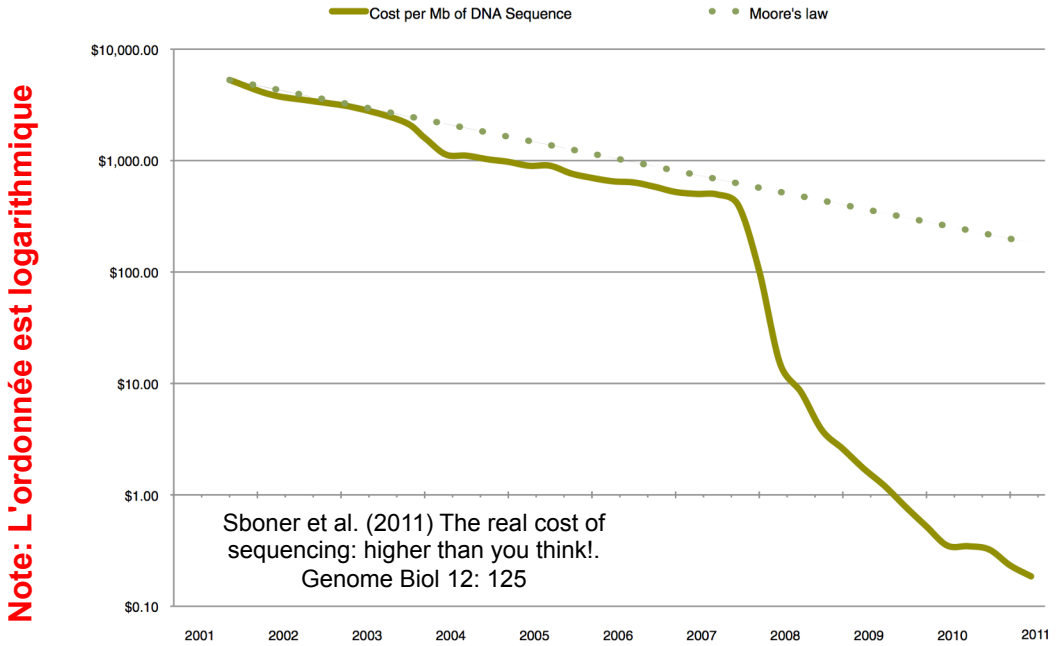


Figure 2. Cost of 1 MB of DNA sequencing. Decreasing cost of sequencing in the past 10 years compared with the expectation if it had followed Moore's law. Adapted from [11]. Cost was calculated in January of each year. MB, megabyte.

- Figure from: Sboner et al. The real cost of sequencing: higher than you think!. Genome Biol (2011) vol. 12 (8) pp. 125

Le vrai coût des projets de séquençage

- La chute des prix du séquençage va de pair avec une augmentation des coûts relatifs d'autres étapes du projet:
  - Pre-processing: collection et préparation des échantillons.
  - Post-processing: analyse des données massives générées par les projets.
- Les laboratoires qui se lancent dans le séquençage à haut débit expriment donc un besoin croissant pour l'analyse bioinformatique.

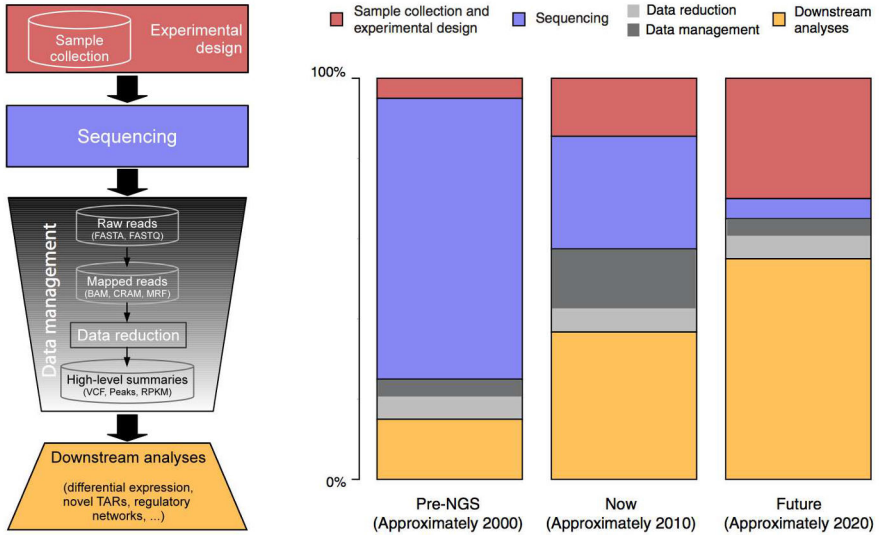


Figure 1. Contribution of different factors to the overall cost of a sequencing project across time. Left, the four-step process: (i) experimental design and sample collection, (ii) sequencing, (iii) data reduction and management, and (iv) downstream analysis. Right, the changes over time of relative impact of these four components of a sequencing experiment. BAM, Binary Sequence Alignment/Map; BED, Browser Extensible Data; CRAM, compression algorithm; MRF, Mapped Read Format; NGS, next-generation sequencing; TAR, transcriptionally active region; VCF, Variant Call Format.

Sboner et al. (2011) The real cost of sequencing: higher than you think!. Genome Biol 12: 125

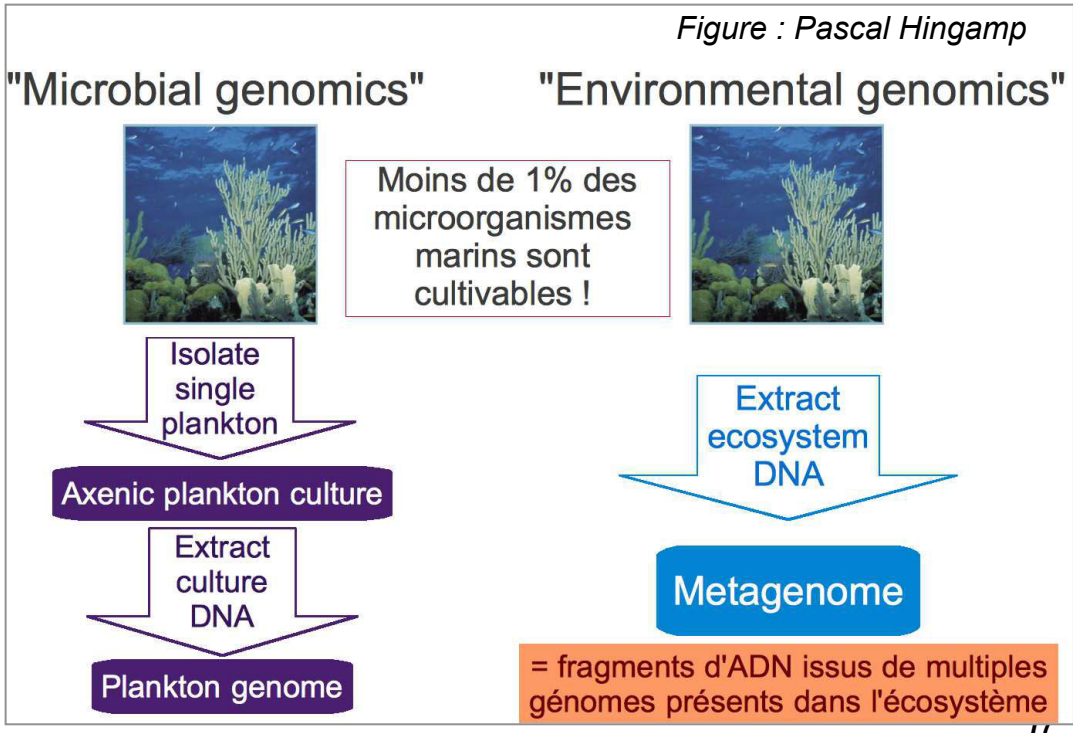
- Sboner et al. (2011) The real cost of sequencing: higher than you think!. Genome Biol 12: 125

Métagénomique - échantillonnage des génomes

- La métagénomique consiste à séquencer des échantillons provenant de divers milieux (océans, flore intestinale, …) pour échantillonner les espèces vivantes dans leur milieu naturel.
- Exemple: l'expédition TARA a échantillonné de la biodiversité dans les eaux océaniques de 2010 à 2012. L'analyse de ces échantillons poursuit son cours.
- En approche « génomique classique », on isole une espèce microbienne, on la met en culture, et on séquence ensuite son génome (si la culture fonctionne).
- En approche métagénomique, on séquence directement tout l'ADN extrait de l'écosystème.
- On peut ensuite
  - identifier les espèces présentes,
  - caractériser leur abondance,
  - découvrir de nouvelles protéines,
  - ....

http://oceans.taraexpeditions.org/

Figure : Pascal Hingamp



Etudes d'associations à l'échelle du génome complet

- La technologie des biopuces permet de caractériser à échelle génomique les variations interindividuelles.
- Une étude a été menée sur 17.000 personnes afin d'identifier les régions génomiques associées à 7 maladies (2.000 patients par maladie) par rapport à un groupe de contrôle (3.000 personnes).
- La figure synthétise les résultats, en indiquant (en vert) les SNPs associés de façon significative à l'une des maladies.
  - Les zones bleues représentent les chromosomes.
  - Chaque point vert représente un SNP, et sa la hauteur indique la significativité.

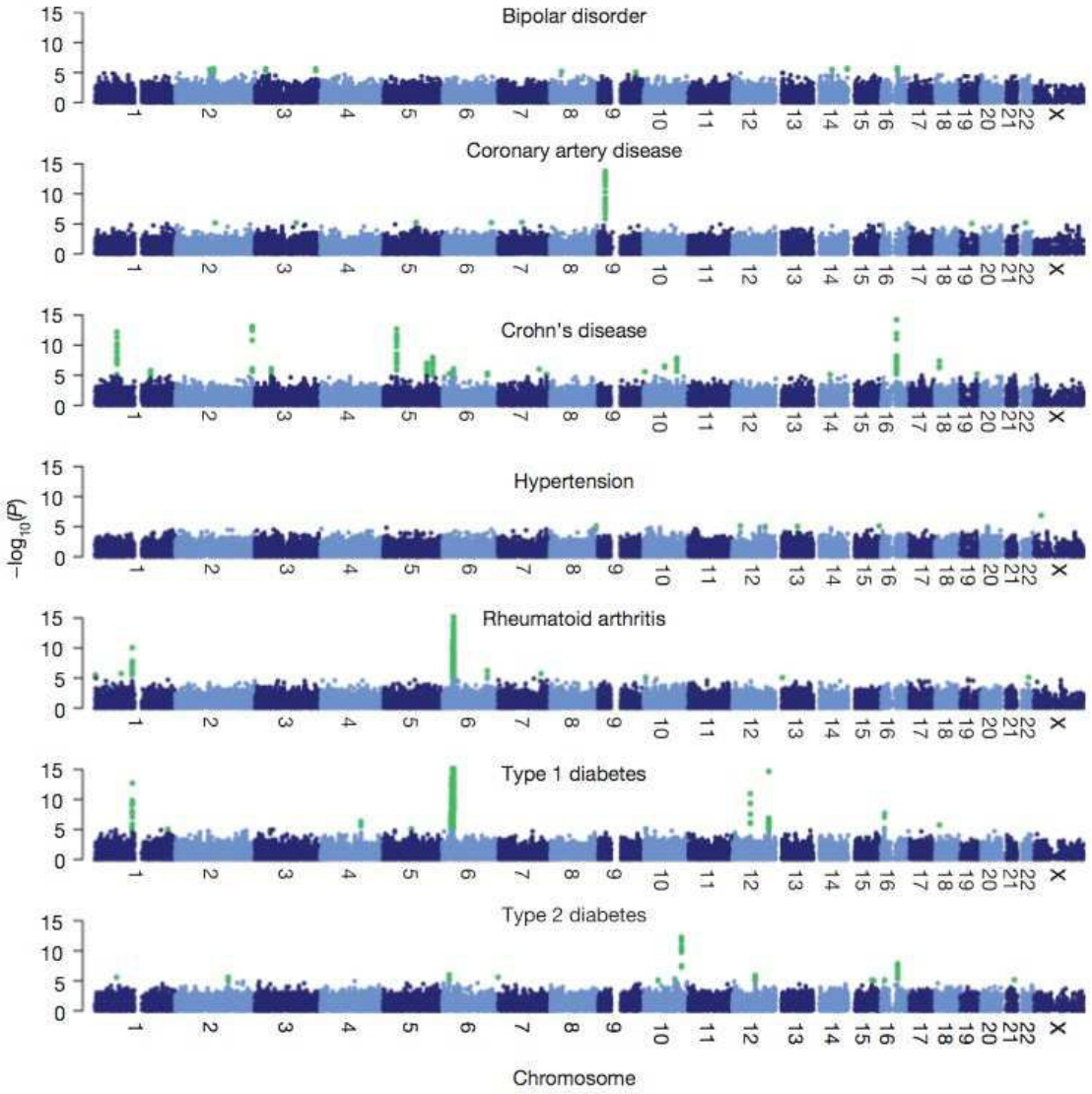
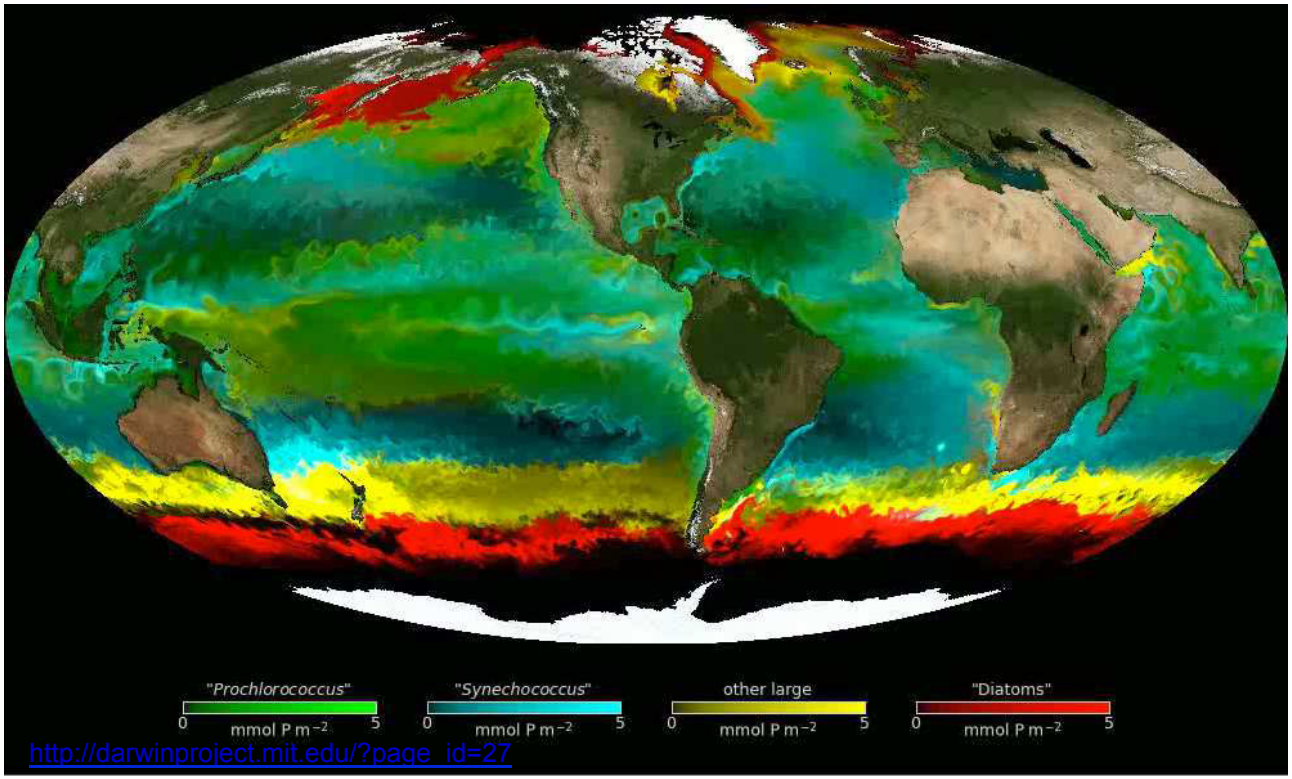


Figure 4 | Genome-wide scan for seven diseases. For each of seven diseases $-\log_{10}$ of the trend test P value for quality-control-positive SNPs, excluding those in each disease that were excluded for having poor clustering after visual inspection, are plotted against position on each chromosome. Chromosomes are shown in alternating colours for clarity, with P values $< 1 \times 10^{-5}$ highlighted in green. All panels are truncated at $-\log_{10}(P \text{ value}) = 15$, although some markers (for example, in the MHC in T1D and RA) exceed this significance threshold.

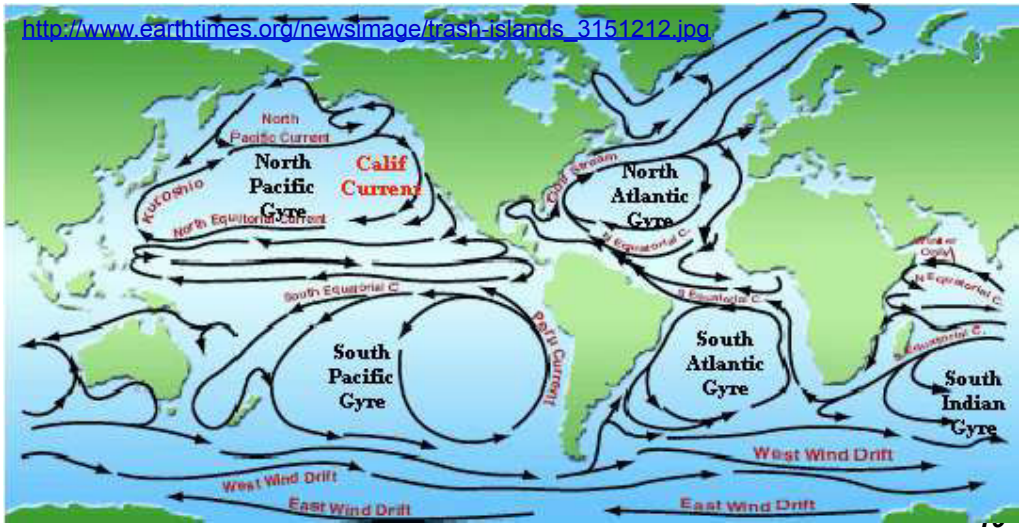
- Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661-78.

Les fluctuations dynamiques des espèces planctoniques

- Haut: modèle dynamique (animation) de la densité d'espèces planctoniques pendant 7 cycles annuels. Résultat du projet TARA aimablement fourni par Pascal Hinamp.
- Bas: courants résultant des forces de Coriolis.

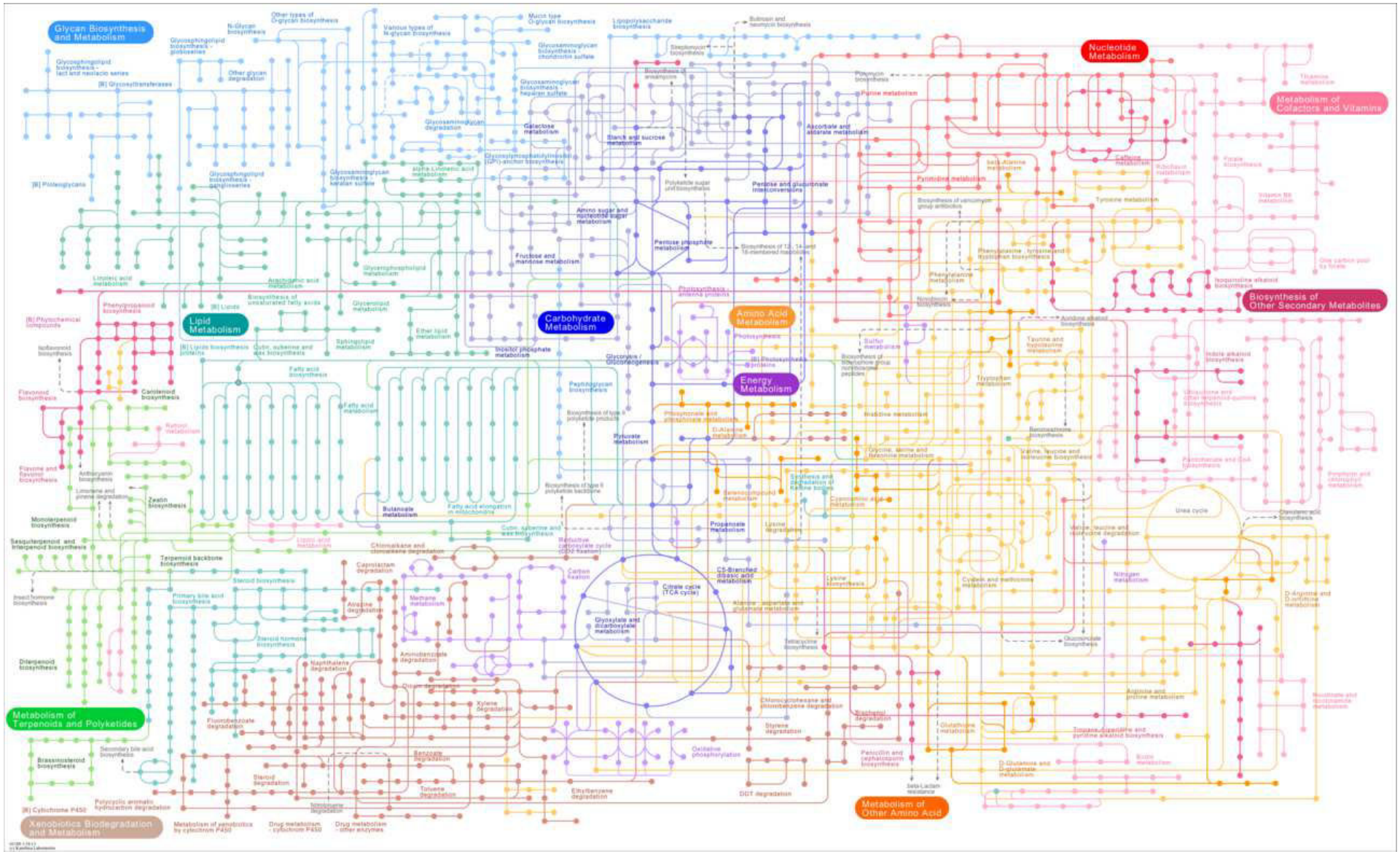


http://darwinproject.mit.edu/?page_id=27



http://www.earthtimes.org/newsimage/trash-islands_3151212.jpg

Analysis of metabolic networks



Phylogénomique

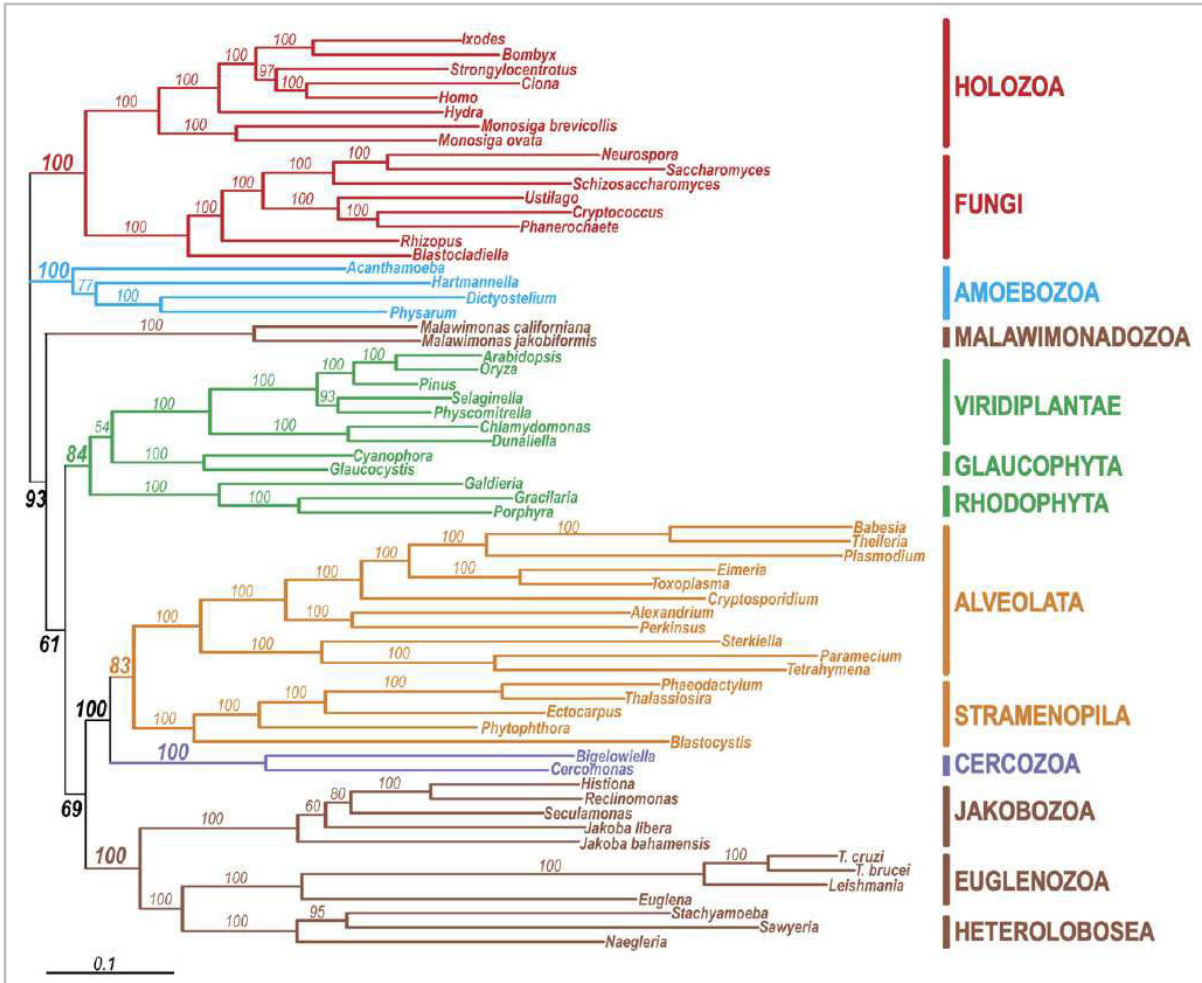


Figure 1. Maximum-Likelihood Tree of Eukaryotes

The tree includes 64 species and is based on 143 concatenated nucleus-encoded proteins (31,604 amino acid positions). Numbers indicate support values of RaxML analysis (100 replicates) with the WAG + F + Γ model. Posterior probabilities obtained in the Bayesian Inference with MrBayes are 1.0 for all branches. The scale bar denotes the estimated number of amino acid substitutions per site. The tree was rooted according to a gene fusion [13, 16].

- En phylogénie moléculaire, une approche classique consiste à se concentrer sur un gène considéré comme représentatif, et à construire un arbre sur base de la divergence de séquence de ce gène.
- Ces approches peuvent maintenant être généralisées en comparant les séquences de plusieurs centaines de gènes.
- Elles permettent d'inférer des phylogénies entre organismes très éloignes (règnes différents), et d'établir ainsi des scénarios concernant les premières étapes de la diversification des êtres vivants.

- Source: Rodríguez-Ezpeleta et al. Curr Biol (2007) vol. 17 (16) pp. 1420-5
- Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans.

L'arbre universel de la vie revisité

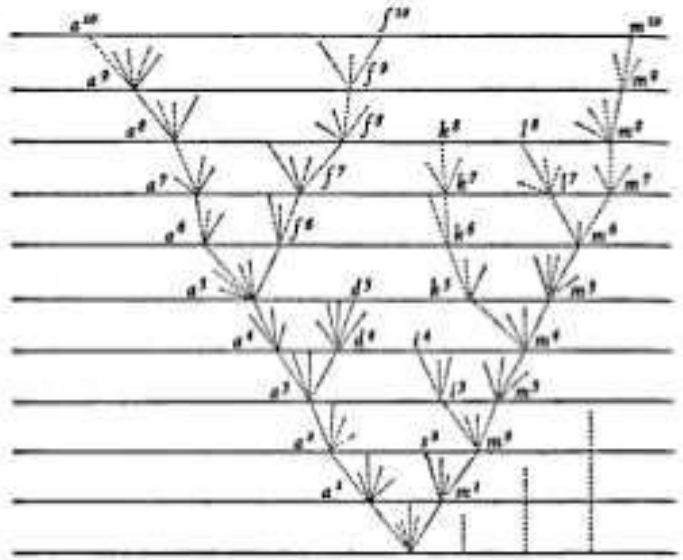


Fig. 1. Part of the only figure in the *Origin of Species*. Darwin first uses it to represent the divergence of variants within a species, showing successively more difference in a single lineage (a^1 through a^{10}) and splitting into multiple lineages (*m*, *s*, *i*, and so forth), some of which will become new species. Later, he expands the tree metaphor, explaining that "limbs divided into great branches ... were themselves once, when the tree was small, budding twigs; and this connection of the former and present buds by ramifying branches may well represent the classification of all extinct and living species in groups subordinate to groups" (*3*, p. 171).

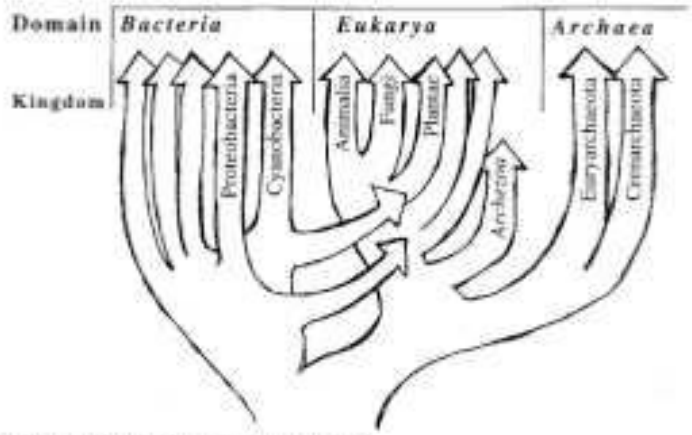


Fig. 2. The current consensus or standard model. Only a few of the "kingdoms" of the "domain" Bacteria are shown. Branching orders of several kingdoms within Bacteria and Eukarya remain in dispute. Mitochondrial and chloroplast endosymbioses are indicated by lower and upper diagonal arrows, respectively. Archezoa, as a subkingdom composed of primitively amitochondriate protists, may be extinct. For SSU rRNA trees with much more detail, see (*5*).

- L'arbre de la vie de Darwin (Fig 1) est revisité par Doolittle (1999) pour tenir compte
  - Fig 2: des événements d'endosymbiose liés à l'apparition des organelles des eucaryotes (mitochondrie et chloroplaste).
  - Fig 3: des transferts horizontaux entre génomes de procaryotes.

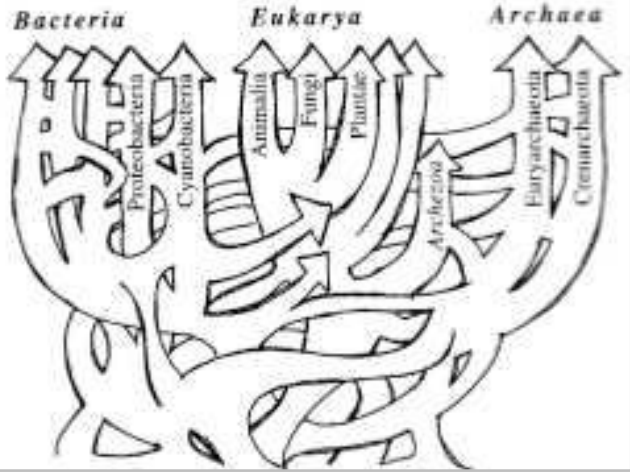


Fig. 3. A reticulated tree, or net, which might more appropriately represent life's history. Martin (*16*), in a review covering many of the same topics as this one, has presented some striking colored representations of such patterns.

Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. Science 284, 2124-9.

L'anneau de la vie

The ring of life provides evidence for a genome fusion origin of eukaryotes

Maria C. Rivera[1,3,4] & James A. Lake[1,2,4]

[1]Molecular Biology Institute, MCD Biology, [2]Human Genetics, [3]IGPP, and [4]Astrobiology Institute, University of California, Los Angeles 90095, USA

Genomes hold within them the record of the evolution of life on Earth. But genome fusions and horizontal gene transfer seem to have obscured sufficiently the gene sequence record such that it is difficult to reconstruct the phylogenetic tree of life. Here we determine the general outline of the tree using complete genome data from representative prokaryotes and eukaryotes and a new genome analysis method that makes it possible to reconstruct ancient genome fusions and phylogenetic trees. Our analyses indicate that the eukaryotic genome resulted from a fusion of two diverse prokaryotic genomes, and therefore at the deepest levels linking prokaryotes and eukaryotes, the tree of life is actually a ring of life. One fusion partner branches from deep within an ancient photosynthetic clade, and the other is related to the archaeal prokaryotes. The eubacterial organism is either a proteobacterium, or a member of a larger photosynthetic clade that includes the Cyanobacteria and the Proteobacteria.

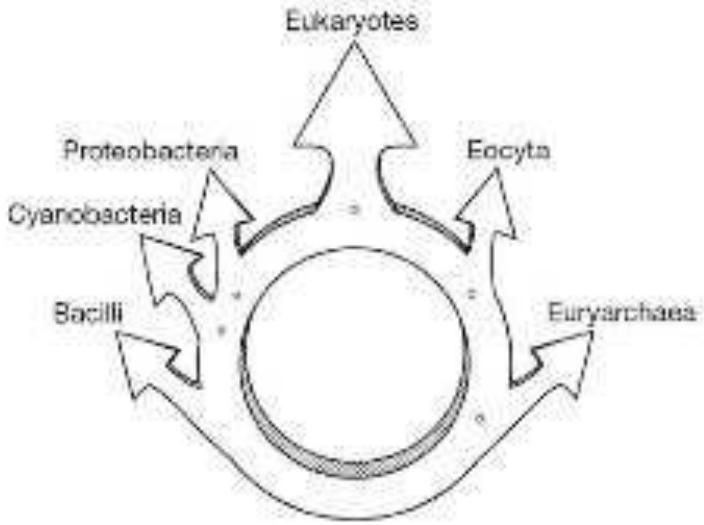
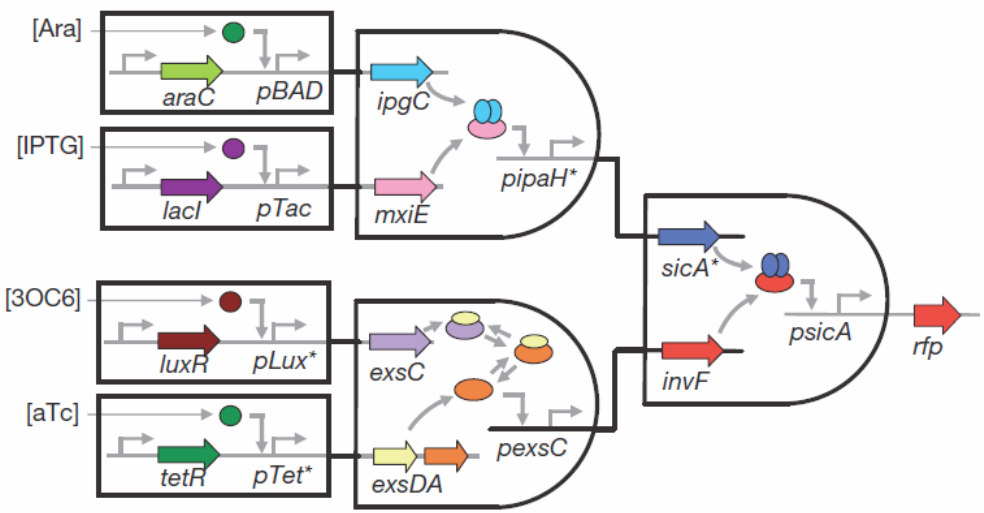


Figure 3 A schematic diagram of the ring of life. The eukaryotes plus the two eukaryotic host organisms (the operational and informational ancestors) comprise the eukaryotic realm (see Supplementary Discussion). Ancestors defining major groups in the prokaryotic realm are indicated by small circles on the ring. The Archaea[44], shown on the bottom right, includes the Euryarchaea, the Eocyta and the informational eukaryotic ancestor. The Karyota[2], shown on the upper right of the ring, includes the Eocyta and the informational eukaryotic ancestor. The upper left circle includes the Proteobacteria[45] and the operational eukaryotic ancestor. The most basal node on the left represents the photosynthetic prokaryotes and the operational eukaryotic ancestor.

- Rivera & Lake (2004) analysent les relations entre tous les gènes d'eukaryotes, d'eubactéries, et d'archées.
- Leur analyse suggère que les génomes eukaryotes résulteraient d'une fusion entre un génome de bactérie et un génome d'archée.
- Les gènes provenant des archées sont majoritairement impliqués dans des fonctions de maintien de la cellule (réplication, transcription et sa régulation).
- Les gènes provenant des archées sont majoritairement impliqués dans le métabolisme.

Rivera, M. C. and Lake, J. A. (2004). The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature 431, 152-5.

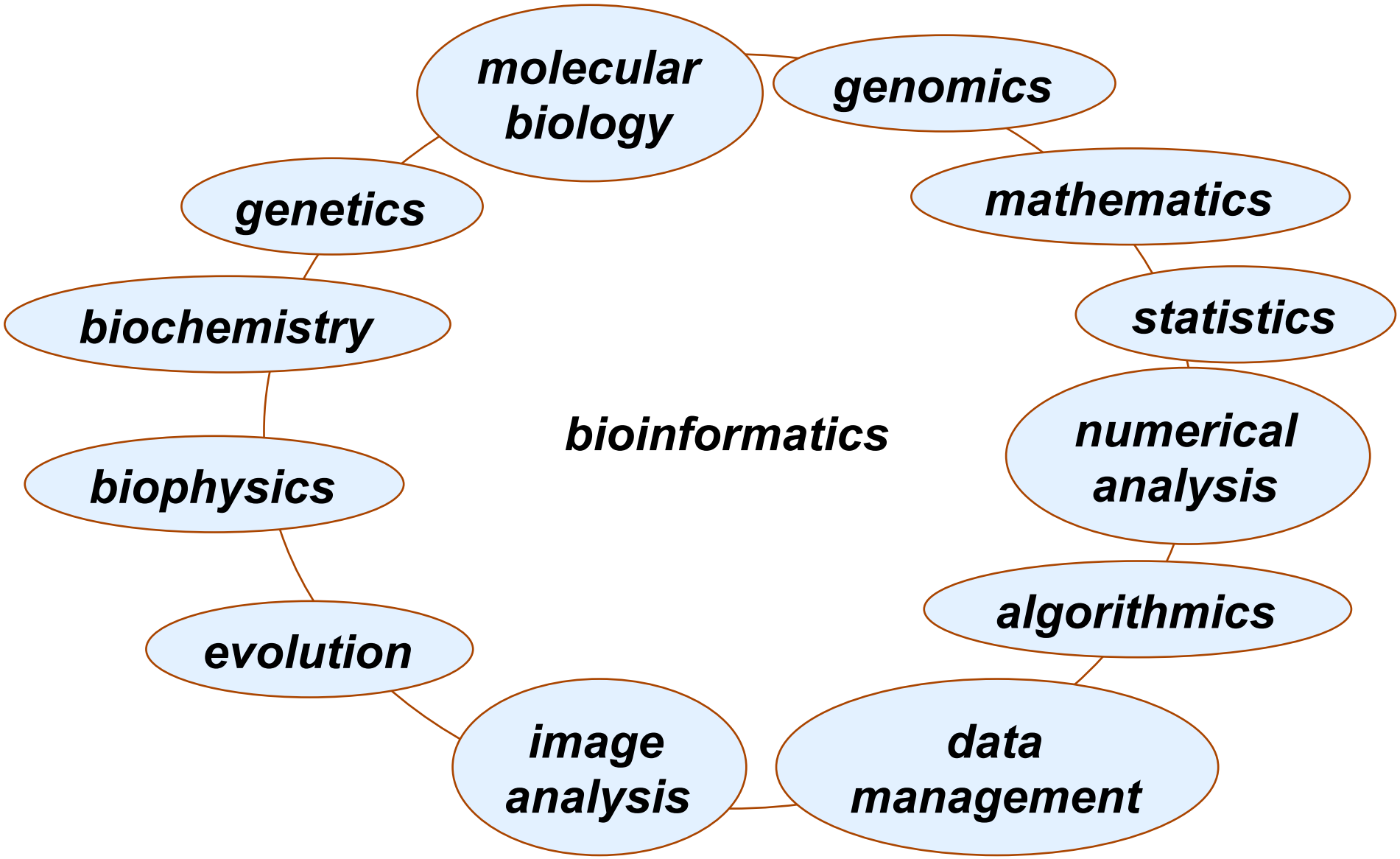
Synthetic biology



http://www.kurzweilai.net/the-most-complex-synthetic-biology-circuit-yet

Bioinformatics – a fast evolving domain

Multidisciplinarity



Multidisciplinarité

- La bioinformatique est un domaine intrinsèquement multidisciplinaire
- Les scientifiques ne peuvent pas être des experts dans tous ces domaines
- Solution: des équipes multidisciplinaires et / ou projets multi-laboratoire
- Problèmes
  - Les biologistes (en général) détestent les statistiques et les ordinateurs
  - Les informaticiens (en général) ne tiennent pas compte des statistiques et de la biologie
  - Statisticiens et mathématiciens (caricature)
    - Parlent une langue étrange pour tout autre être humain.
    - Passent leur temps à écrire des formules incompréhensibles
  - Complexité du domaine biologique
    - Chaque fois que vous essayez de formuler une règle, il y a un contre-exemple
    - Même la définition d'un mot unique requiert un livre plutôt que d'une phrase (exercice: trouver une définition consensuelle du «gène»)

Formations en bioinformatique

- Le problème de l'interdisciplinarité
  - L'interdisciplinarité nécessite de la communication.
  - La communication nécessite un vocabulaire partagé
- Diversité des objectifs de formations en bioinformatique
  - Former les étudiants en biologie à l'utilisation de méthodes bioinformatiques utilisées pour l'analyse des données biologiques
  - Former les informaticiens, mathématiciens à comprendre les données biologiques, afin de développer de nouvelles méthodes analytiques.
  - Former des scientifiques multidisciplinaires capables de concevoir de nouvelles approches et de développer de nouveaux outils (devenir bioinformaticiens)
- Typologie des formations en bioinformatique
  - Cours d'introduction pour les biologistes/médecins
  - Formations intensives et courtes (1 semaine) destinées aux chercheurs en biologie/médecine
  - Master en bioinformatique/génomique (1 ou 2 ans)
  - Formation complète en bioinformatique (en Allemagne, Mexique, …) dès la sortie du secondaire.

Exemples d'applications

- Recherche en biologie
  - Organisation moléculaire de la cellule / organisme
  - Biologie du développement
  - Mécanismes de l'évolution
  - Médecine
  - Diagnostic de cancers
  - Détection des gènes impliqués dans le cancer
- La recherche pharmaceutique
  - Mécanismes d'action des médicaments
  - Identification de cibles pharmaceutiques
- Biotechnologie
  - Thérapie génique
  - Bioingénierie
  - Biologie synthétique

From wet science to bioinformatics

- Progresses in biology/biophysics stimulated the incorporation of new methods in bioinformatics
  - Structure analysis (since the 50s)
    - structure comparison
    - structure prediction
  - Sequencing (since the 70s)
    - Sequence alignment
    - Sequence search in databases
  - Genomes (since the 90s)
    - Genome annotation
    - Comparative genomics
    - Functional classifications (“ontologies”)
  - Transcriptome (since 1997)
    - Multivariate analysis
  - Proteome (~ 2000)
    - Network analysis

High throughput technologies

- Genome projects stimulated drastic improvement of sequencing technology
- Post-genomic era
  - Genome sequence was not sufficient to predict gene function
  - This stimulated the development of new experimental methods
    - transcriptomics (microarrays)
    - proteomics (2-hybrid, mass spectrometry, ...)
- Warning: the "omics" trends
  - The few real high throughput methods raised a fashion of "omics", which introduced more confusion than progress
  - Some of the "omics" are not associated to any new/high throughput approach, this is just a new name on a previous method, or on an abstract concept

Large-scale analyses

- The availability of massive amounts of data enables to address questions that could not even be imagined a few years ago
  - genome-scale measurement of transcriptional regulation
  - comparative genomics
- Most of the downstream analyses require a good understanding of statistics
- Warning: the global trends
  - the capability to analyze large amounts of data presents a risk to remain at a superficial level, or to be fooled by forgetting to check the pertinence of the results (with some in-depth examples)
  - good news: this does not prevent the authors from publishing in highly quoted journals

The risks of inference

- Bioinformatics is essentially a science of inference
- Any analysis of massive data will unavoidably generate a certain rate of errors (***false positives*** and ***false negatives***).
- Good research and development will include an evaluation of the error rates.
- Good methods will minimize the error rate.
- However, there is always a trade between specificity and sensitivity.

Why to do bioinformatics then ?

- In most cases, wet biology will be required afterwards to validate the predictions
- Bioinformatics can
  - reduce the universe of possibilities to a small set of testable predictions
  - assign a degree of confidence to each prediction
- The biologist will often have to chose the appropriate degree of confidence, depending on the trade between
  - cost for validating predictions
  - benefit expected from the right predictions
- Bioinformatics as *in silico* biology
  - Beyond its role in generating testable hypotheses for the biologist, bioinformatics also allows to explore domains that can not be addressed experimentally.
  - A typical example is the study of past evolutionary events
    - Phylogenetic inference and comparative genomics give us insights in the mechanisms of evolution and in the past evolutionary events
    - The time scale of these events is however so large (billions of years) that one cannot conceive to reproduce the inferred events with experimental methods.

Books

- Zvelebil, M. & Baum, J.O. Understanding Bioinformatics. (2007) pp. 772
- Pevzner, J. (2003). Bioinformatics and Functional Genomics. Wiley.
  - All the slides available at: http://www.bioinfbook.org/
- W. Mount. Bioinformatics: Sequence and Genome Analysis. (2004) pp. 692.
  - http://www.bioinformaticsonline.org/
- Westhead, D.R., J.H. Parish, and R.M. Twyman. 2002. ***Bioinformatics***. BIOS Scientific Publishers, Oxford.
- Branden et al. Introduction to Protein Structure. (1998) pp. 410