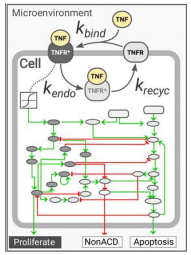
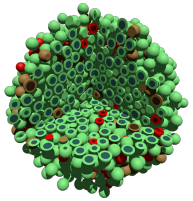
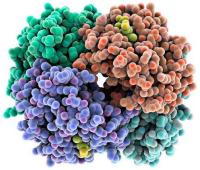
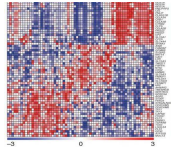




De las redes intracelulares a las simulaciones de sistemas multicelulares





Profesor: Dr. Miguel Ponce de León (miguel.ponce@bs.es) - BSC

Coordinador: Dr. Flavio Pazos (flavio.pazos@gmail.com) - IIBCE/IP

Apoyan:

PEDECIBA MEC-UDELAR

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Institut Pasteur de Montevideo

iibce
INSTITUTO DE INVESTIGACIONES BIOLÓGICAS CLEMENTE ESTABLE

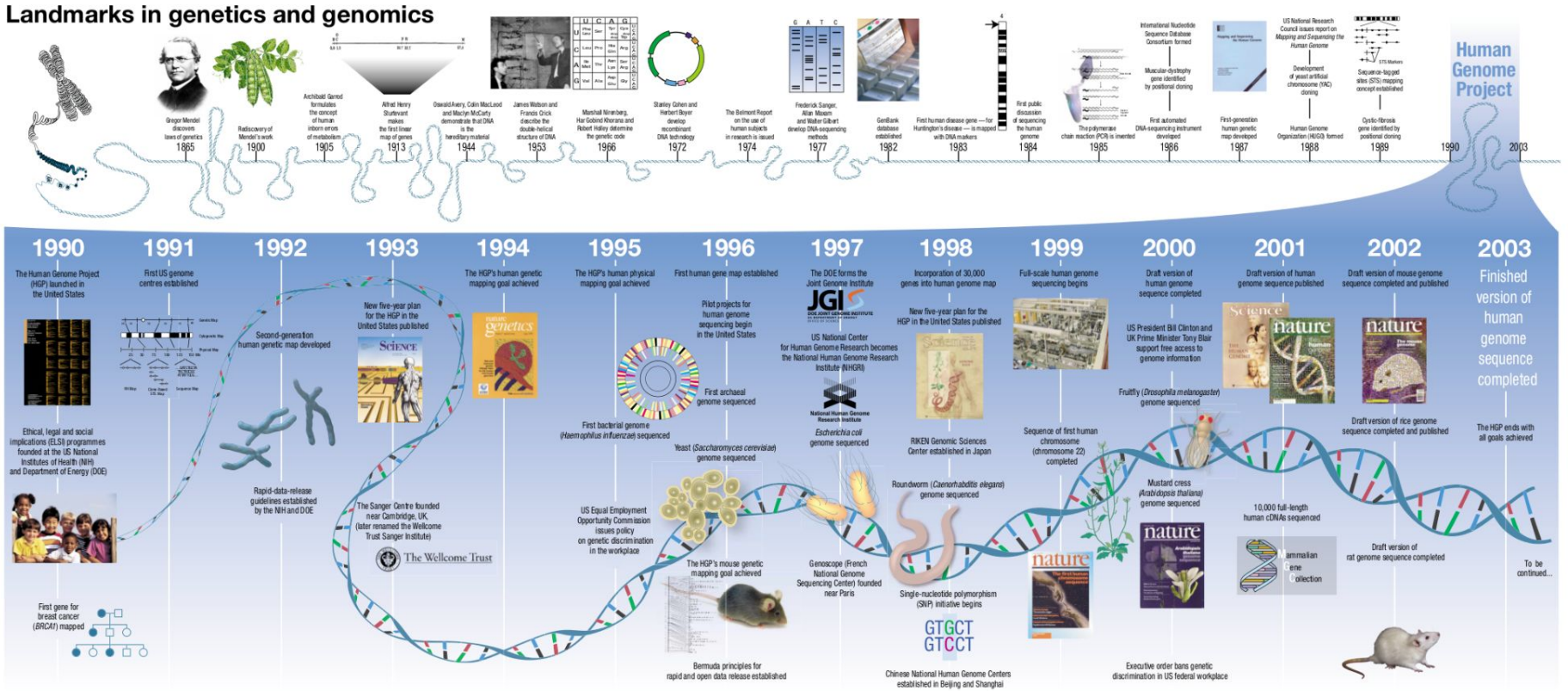
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Tema 2

Anotación genómica y reconstrucción de redes moleculares

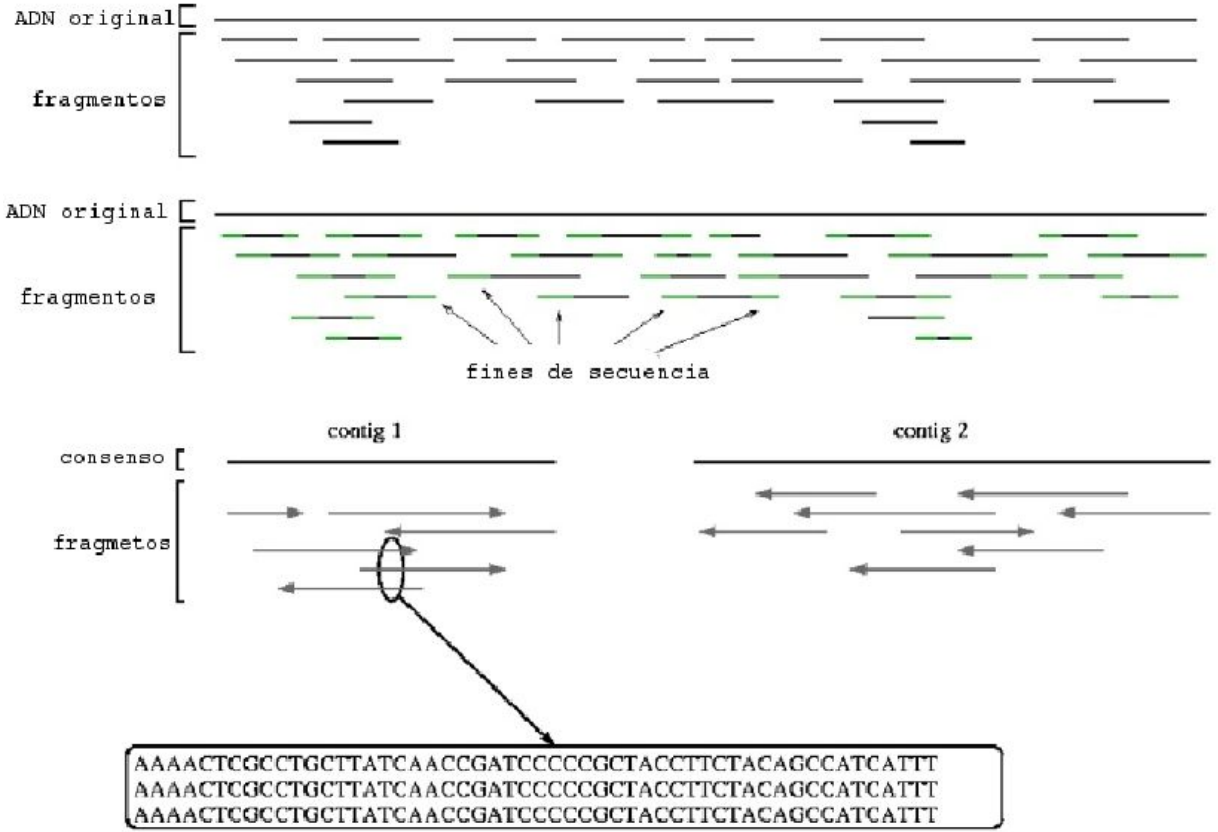
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Nace la primera *omica* ⟶ Genómica



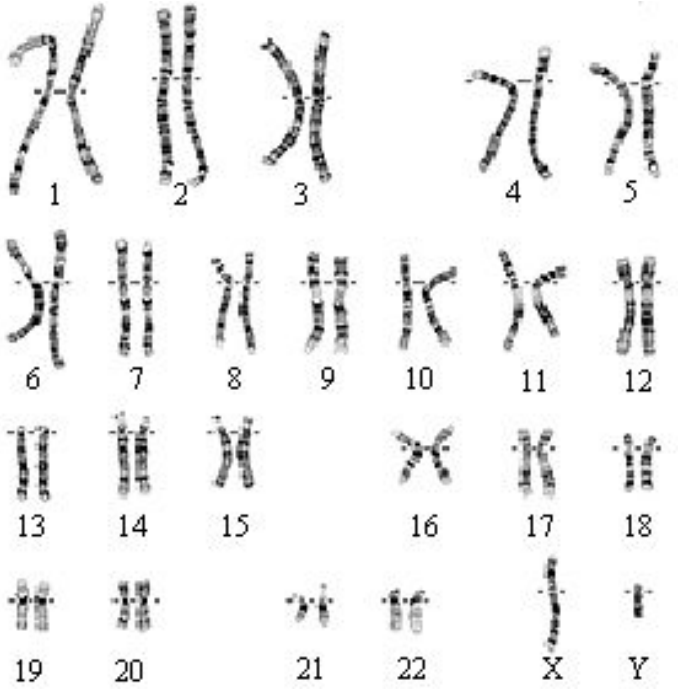
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Secuenciación y ensamblado de genomas → nace la bioinformática

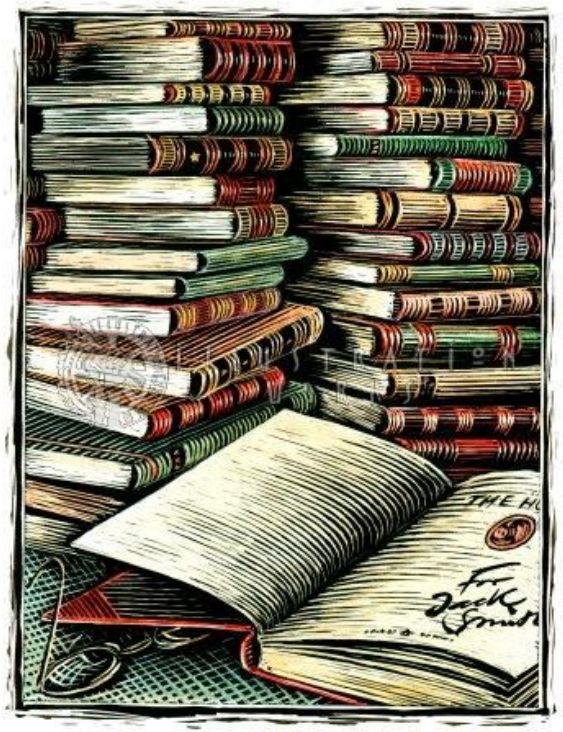


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

El genoma humano en números (o letras)



$\simeq$



3.000.000.000 pares de bases

200 guías telefónicas de Manhattan
(1000 páginas cada una!!!)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

El genoma humano en números (o letras)

3000 pb ~ 0.0001% del genoma humano (0.06% E. *coli*)

TATGCCGACCATTGCGCTGACTAACAGCATCGCTTTTGCCAACGTGCCGGATGTGGAGCGTGATTTCCCGCGCATTCGTGTGATGGGCACTATCGGCTGGATTGCCGCCGGTCTGGCATGTGGT
TTCTTGCCGCAAATACTGGGGTATGCCGATATCTCACCGACTAACATCCCGCTGCTGATTACCGCCGGAAGTTCTGCTCTGCTCGGTGTGTTTGCGTTTTTCCTGCCCGACACGCCACCAAAAA
GCACCGGCAAAATGGATATTAAAGTCATGCTCGGCCTGGATGCGCTGATCCTGCTGCGCGATAAAAACTTCCTCGTCTTTTTCTTCTGTTCATTCCTGTTTGCGATGCCACTGGCGTTCTATTA
CATCTTTGCCAACGGTTATCTGACCGAAGTCGGCATGAAAAATGCCACCGGCTGGATGACGCTCGGCCAGTTCTCTGAAATATTCTTTATGCTGGCATTACCGTTTTTCACTAAACGCTTTGGT
ATCAAAAAGGTATTATTGCTTGGTCTGGTCACCGCTGCGATCCGCTATGGCTTCTTTATTTACGGTAGTGCGGATGAATATTTCACCTACGCGTTACTGTTCCTCGGTATTTTGCTTCACGGCG
TAAGTTACGATTTTTACTACGTTACCGCTTACATCTATGTCGATAAAAAAGCCCCCGTGCATATGCGTACCGCCGCGCAGGGGCTGATCACGCTCTGCTGCCAGGGCTTCGGCAGTTTGCTCGG
CTATCGTCTTGGCGGTGTGATGATGGAAAAGATGTTCGCTTATCAGGAACCGGTAAACGGACTGACTTTCAACTGGTCCGGGATGTGGACTTTTGGCGCGGTGATGATTGCCATTATCGCCGTG
CTGTTCATGATTTTTTTCCGCGAATCCGACAACGAAATTACGGCTATCAAGGTCGATGATCGCGATATTGCGTTGACACAAGGGGAAGTTAAATGAAAACAGAACGTATTCTCGGTGCTCTTTA
TGGGCAGGCGTTAGGGGATGCGATGGGGATGCCCTCCGAGCTTTGGCCACGCAGCCGCGTTAAAGCACACTTTGGCTGGATTGACCGTTTTCTGCCTGGACCAAAGGAGAATAACGCGGCCTGT
TATTTTAACCGCGCCGAATTCACCGACGATACCTCGATGGCGCTGTGTCTGGCGGATGCGTTACTGGAACGTGAAGGCAAGATCGATCCGAATCTGATTGGGCGTAATATTCTCGACTGGGCGC
TGCGTTTCGACGCCTTTAACAAAAACGTACTTGGTCCGACCTCGAAGATTGCGCTTAACGCCATTCGCGACGGTAAACCCATTGCTGAACTGGAAAACAACGGAGTGACCAACGGCGCAGCGAT
GCGCGTCTCGCCATTAGGTTGTTTGCTTCCGGCGCACGATGTTGATTCCTTTATTGATGATGTGGCGCTGGCGTCCAGCCCGACCCATAAATCCGATCTGGCGGTTGCAGGCGCGGTAGTCATC
GCATGGGCGATTTCTCGTGCCATTGACGGAGAAAGCTGGTCAGCGATTGTAGATTCACTGCCTTCAATTGCGCGACATGCACAGCAAAAACGCATCACCACGTTCAGCGCCTCACTGGCATCAC
GTCTGGAGATTGCGCTGAAAATTGTGCGCAATGCCGACGGCACCGAATCCGCCAGCGAACAGCTTTACCAGGTCGTTGGCGCAGGTACCAGCACTATTGAGTCCGTTCCGTGCGCCATTGCGCT
GGTTGAACTGGCACAAACCGACCCGAATCGCTGCGCCGTCCTGTGCGCTAACCTTGGCGGCGACACAGACACCATCGGCGCTATGGCGACGGCGATTTGCGGCGCGTTGCATGGCGTTAACGCT
ATCGATCCTGCATTAAAGGCGGAACTGGATGCGGTAAATCAGCTTGATTTCAACCGCTATGCCACAGCGTTGGCGAAATATCGTCAACAACGGGAGGCGGTATGAGCGGCGCTCGATTACACAC
GCTGCTGCCTGAATTAACCACGCGTCAGCCGGTGATGGTCGTCGGCGCGGCGGTCATTGACGTGATCGCCGACGCTTATGCCCTCCCCTGGCGTGGGTGCGATATCGAACTGAAACAGCAGAGC
GTTAACGTTGGCGGCTGCGCACTGAATATTGCCGTGGCGTTAAAGCGCCTCGGCATCGAAGCGGGTAATGCCTTGCCGCTCGGTCAGGGCGTGTGGGCGGAGATTATTCGCAACCGGATGGCAA
AAGAGGGCTTAATCAGCCTGATCGATAACGCCGAAGGTGATAACGGCTGGTGTCTGGCGCTGGTTGAGCCGGATGGCGAACGCACTTTTATGTCATTCAGCGGTGTGGAAAATCAGTGGAATCG
CCAGTGGCTGGCGCGATTAACCGTTGCGCCTGGCAGCCTGCTCTATTTTTCCGGTTATCAACTGGCCTCGCCCTGCGGCGAATTGTTAGTGGAATGGCTGGAAGAGCTGCAAGACGTGACGCCG
TTTATCGATTTTGGCCCACGTATTGGCGATATCCCGGATGCATTACTGGCGCGGATCATGGCCTGTCGACCTTTAGTGTCGCTCAATCGTCAAGAGGCTGAGATTGCCGCCGAACGTTTTGCTT
TATCCGCAGAGATAACAACACTTGGCGAGCAATGGCAGGAGAAATTTGCCGCGCCGTTGATCGTTCGTCTCGATAAAGAAGGCGCATGGTATTTCAGCAACGACGCTTCTGGCTGCATTCCGGC
ATTTCCAACGCAAGTTGTAGACACCATTGGGGCGGGCGACAGTCATGCCGGTGGCGTACTTGCCGGGCTGGCCTCTGGTCTGCCACTGGCGGATGCCGTATTACTGGGCAATGCAGTGGCGTCG
TGGGTTGTCGGGCATCGGGGCGGTGATTGTGCGCCAACGCGCGAGGAACTACTCCTCGCACACAAAAACGTATAGATCGCTGCGACAGTGGCTAATGCTGTACTCAATAGGCCGTTGCTGTTGG
TCAAGCGCCACTTGCTTGATCACC

¿Dónde están los genes?

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

El genoma humano en números (o letras)

3000 pb ~ 0.0001% del genoma humano (0.06% E. *coli*)

TATGCCGACCATTGCGCTGACTAACAGCATCGCTTTTGCCAACGTGCCGGATGTGGAGCGTGATTTCCCGCGCATTCGTGTGATGGGCACTATCGGCTGGATTGCCGCCGGTCTGGCATGTGGT
TTCTTGCCGCAAATACTGGGGTATGCCGATATCTCACCGACTAACATCCCGCTGCTGATTACCGCCGGAAGTTCTGCTCTGCTCGGTGTGTTTGCGTTTTTCCTGCCCGACACGCCACCAAAAA
GCACCGGCAAAATGGATATTAAAGTCATGCTCGGCCTGGATGCGCTGATCCTGCTGCGCGATAAAAACTTCCTCGTCTTTTTCTTCTGTTCATTCCTGTTTGCGATGCCACTGGCGTTCTATTA
CATCTTTGCCAACGGTTATCTGACCGAAGTCGGCATGAAAAATGCCACCGGCTGGATGACGCTCGGCCAGTTCTCTGAAATATTCTTTATGCTGGCATTACCGTTTTTCACTAAACGCTTTGGT
ATCAAAAAGGTATTATTGCTTGGTCTGGTCACCGCTGCGATCCGCTATGGCTTCTTTATTTACGGTAGTGCGGATGAATATTTCACCTACGCGTTACTGTTCCTCGGTATTTTGCTTCACGGCG
TAAGTTACGATTTTTACTACGTTACCGCTTACATCTATGTCGATAAAAAAGCCCCCGTGCATATGCGTACCGCCGCGCAGGGGCTGATCACGCTCTGCTGCCAGGGCTTCGGCAGTTTGCTCGG
CTATCGTCTTGGCGGTGTGATGATGGAAAAGATGTTCGCTTATCAGGAACCGGTAAACGGACTGACTTTCAACTGGTCCGGGATGTGGACTTTTGGCGCGGTGATGATTGCCATTATCGCCGTG
CTGTTCATGATTTTTTTCCGCGAATCCGACAACGAAATTACGGCTATCAAGGTCGATGATCGCGATATTGCGTTGACACAAGGGGAAGTTAAATGAAAACAGAACGTATTCTCGGTGCTCTTTA
TGGGCAGGCGTTAGGGGATGCGATGGGGATGCCCTCCGAGCTTTGGCCACGCAGCCGCGTTAAAGCACACTTTGGCTGGATTGACCGTTTTCTGCCTGGACCAAAGGAGAATAACGCGGCCTGT
TATTTTAACCGCGCCGAATTCACCGACGATACCTCGATGGCGCTGTGTCTGGCGGATGCGTTACTGGAACGTGAAGGCAAGATCGATCCGAATCTGATTGGGCGTAATATTCTCGACTGGGCGC
TGCGTTTCGACGCCTTTAACAAAAACGTACTTGGTCCGACCTCGAAGATTGCGCTTAACGCCATTCGCGACGGTAAACCCATTGCTGAACTGGAAAACAACGGAGTGACCAACGGCGCAGCGAT
GCGCGTCTCGCCATTAGGTTGTTTGCTTCCGGCGCACGATGTTGATTCCTTTATTGATGATGTGGCGCTGGCGTCCAGCCCGACCCATAAATCCGATCTGGCGGTTGCAGGCGCGGTAGTCATC
GCATGGGCGATTTCTCGTGCCATTGACGGAGAAAGCTGGTCAGCGATTGTAGATTCACTGCCTTCAATTGCGCGACATGCACAGCAAAAACGCATCACCACGTTCAGCGCCTCACTGGCATCAC
GTCTGGAGATTGCGCTGAAAATTGTGCGCAATGCCGACGGCACCGAATCCGCCAGCGAACAGCTTTACCAGGTCGTTGGCGCAGGTACCAGCACTATTGAGTCCGTTCCGTGCGCCATTGCGCT
GGTTGAACTGGCACAAACCGACCCGAATCGCTGCGCCGTCCTGTGCGCTAACCTTGGCGGCGACACAGACACCATCGGCGCTATGGCGACGGCGATTTGCGGCGCGTTGCATGGCGTTAACGCT
ATCGATCCTGCATTAAAGGCGGAACTGGATGCGGTAAATCAGCTTGATTTCAACCGCTATGCCACAGCGTTGGCGAAATATCGTCAACAACGGGAGGCGGTATGAGCGGCGCTCGATTACACAC
GCTGCTGCCTGAATTAACCACGCGTCAGCCGGTGATGGTCGTCGGCGCGGCGGTCATTGACGTGATCGCCGACGCTTATGCCCTCCCCTGGCGTGGGTGCGATATCGAACTGAAACAGCAGAGC
GTTAACGTTGGCGGCTGCGCACTGAATATTGCCGTGGCGTTAAAGCGCCTCGGCATCGAAGCGGGTAATGCCTTGCCGCTCGGTCAGGGCGTGTGGGCGGAGATTATTCGCAACCGGATGGCAA
AAGAGGGCTTAATCAGCCTGATCGATAACGCCGAAGGTGATAACGGCTGGTGTCTGGCGCTGGTTGAGCCGGATGGCGAACGCACTTTTATGTCATTCAGCGGTGTGGAAAATCAGTGGAATCG
CCAGTGGCTGGCGCGATTAACCGTTGCGCCTGGCAGCCTGCTCTATTTTTCCGGTTATCAACTGGCCTCGCCCTGCGGCGAATTGTTAGTGGAATGGCTGGAAGAGCTGCAAGACGTGACGCCG
TTTATCGATTTTGGCCCACGTATTGGCGATATCCCGGATGCATTACTGGCGCGGATCATGGCCTGTCGACCTTTAGTGTCGCTCAATCGTCAAGAGGCTGAGATTGCCGCCGAACGTTTTGCTT
TATCCGCAGAGATAACAACACTTGGCGAGCAATGGCAGGAGAAATTTGCCGCGCCGTTGATCGTTCGTCTCGATAAAGAAGGCGCATGGTATTTCAGCAACGACGCTTCTGGCTGCATTCCGGC
ATTTCCAACGCAAGTTGTAGACACCATTGGGGCGGGCGACAGTCATGCCGGTGGCGTACTTGCCGGGCTGGCCTCTGGTCTGCCACTGGCGGATGCCGTATTACTGGGCAATGCAGTGGCGTCG
TGGGTTGTCGGGCATCGGGGCGGTGATTGTGCGCCAACGCGCGAGGAACTACTCCTCGCACACAAAAACGTATAGATCGCTGCGACAGTGGCTAATGCTGTACTCAATAGGCCGTTGCTGTTGG
TCAAGCGCCACTTGCTTGATCACC

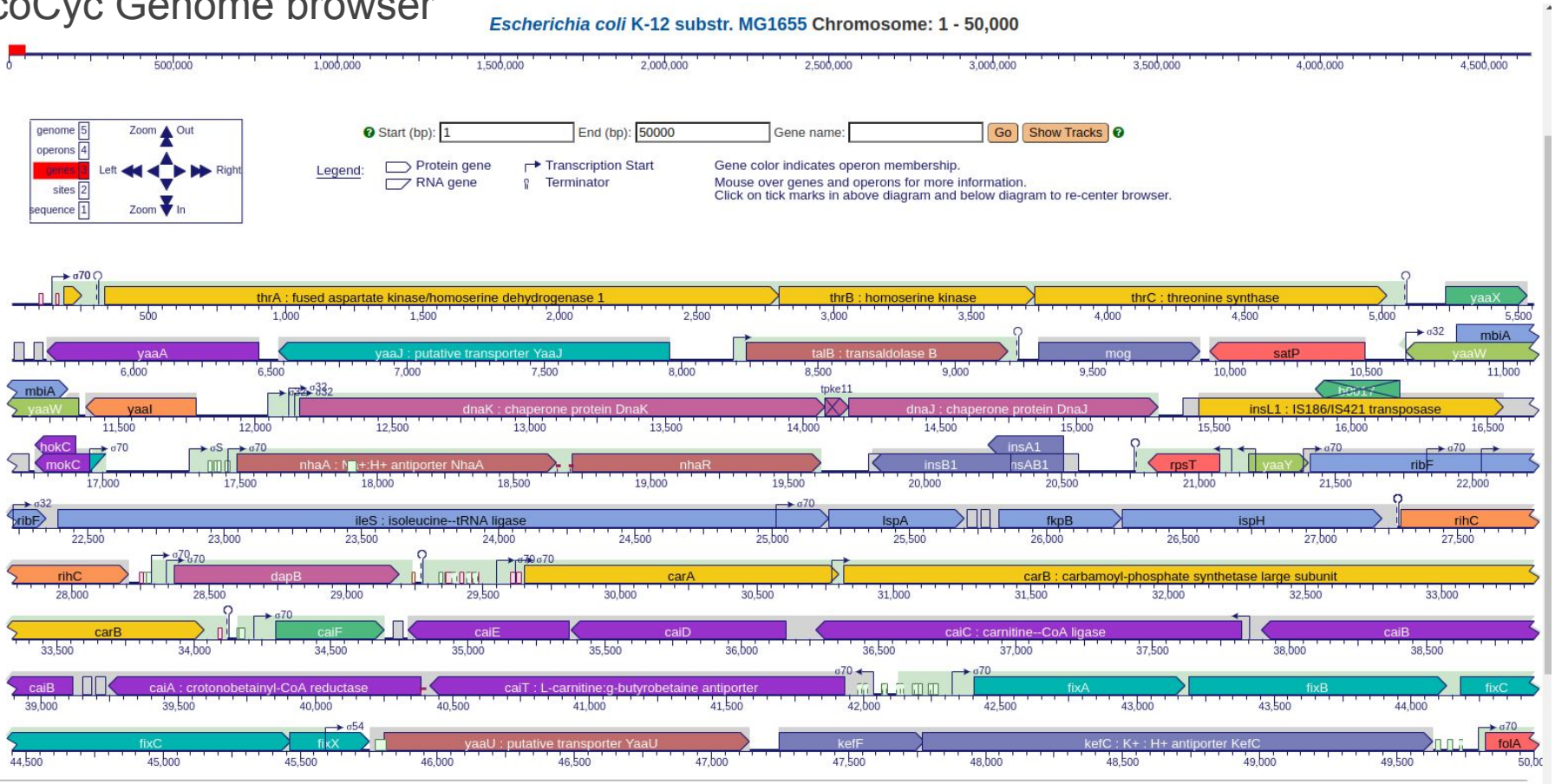
¿Dónde están los genes?

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

¿Cómo se llega a esto?

https://ecocyc.org/

EcoCyc Genome browser

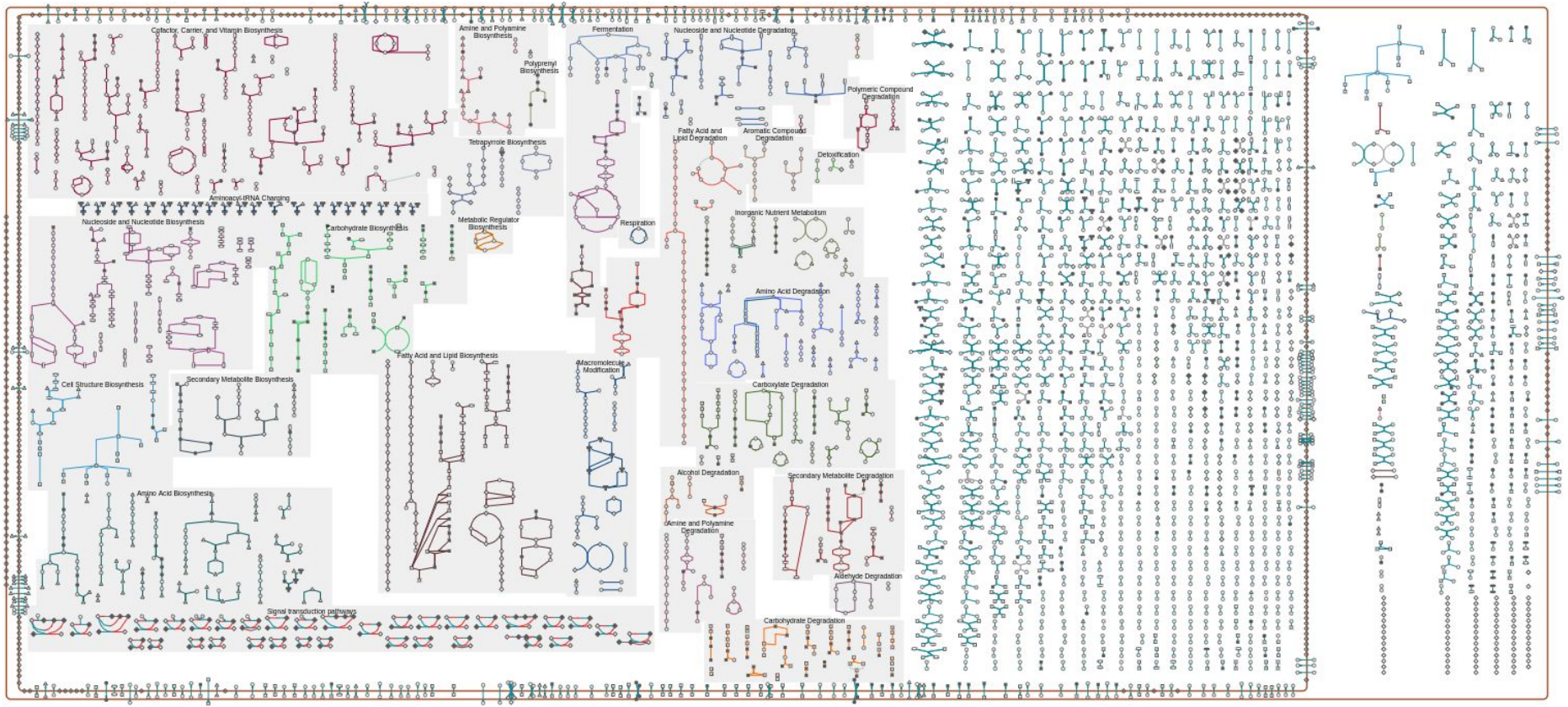


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

¿Cómo se llega a esto?

https://ecocyc.org/

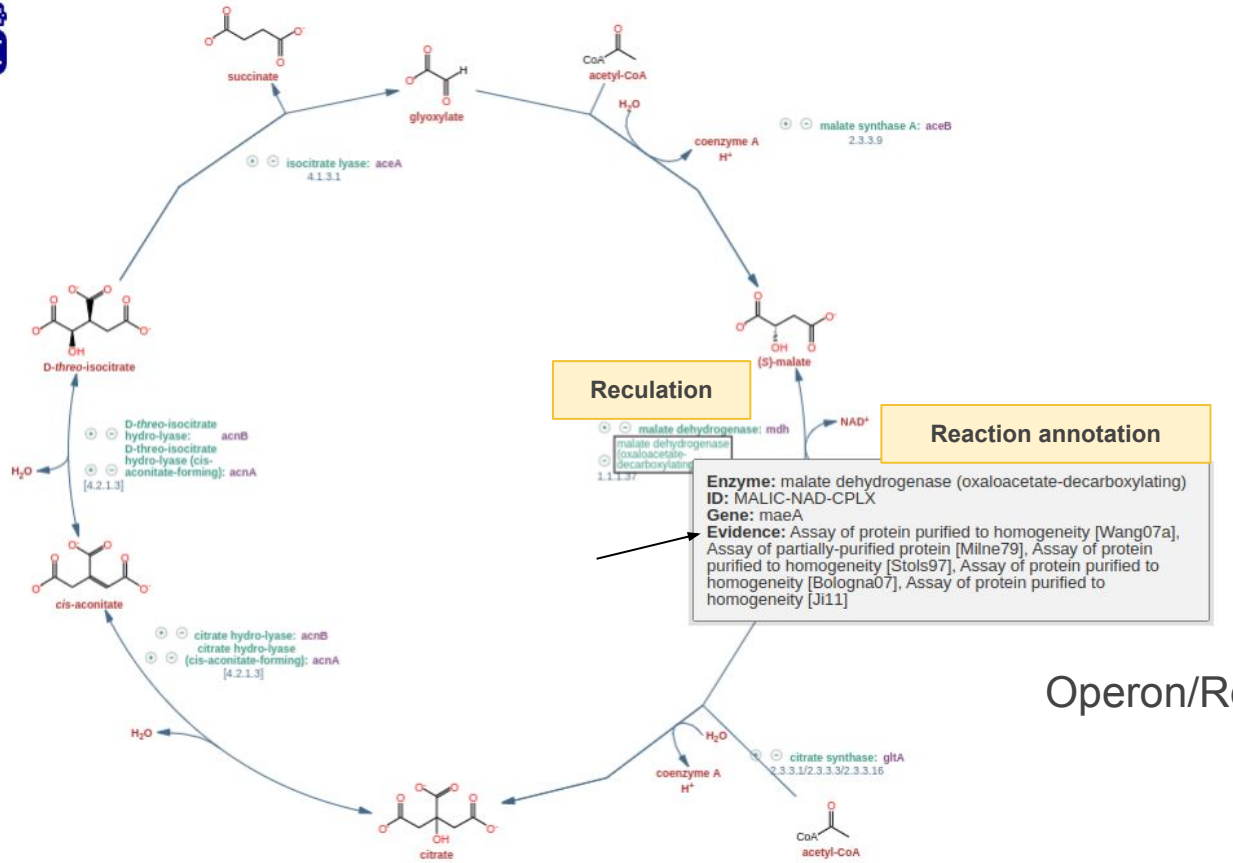
Metabolic reconstruction of E. coli K12 (MG1655). Metabolic browser (EcoCyc)



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

¿Cómo se llega a esto?

https://ecocyc.org/

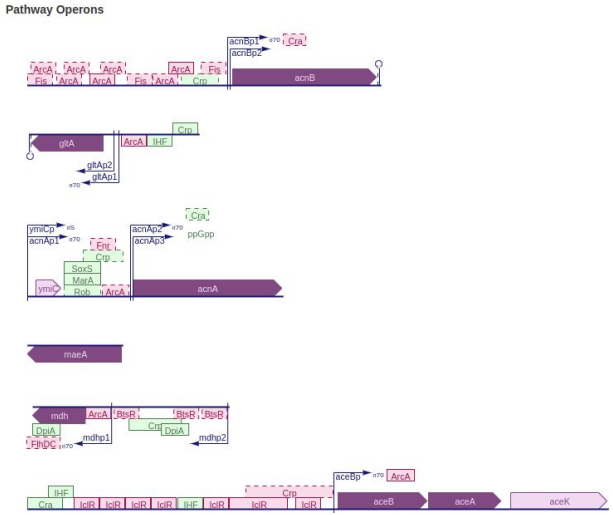


Krebs cycle

Reculation

Reaction annotation

Enzyme: malate dehydrogenase (oxaloacetate-decarboxylating)
ID: MALIC-NAD-CPLX
Gene: maeA
Evidence: Assay of protein purified to homogeneity [Wang07a], Assay of partially-purified protein [Milne79], Assay of protein purified to homogeneity [Stols97], Assay of protein purified to homogeneity [Bologna07], Assay of protein purified to homogeneity [Ji11]



Operon/Regulon structure for the pathway

E. coli K12 (MG1655) Pathway browser (EcoCyc)

GENOME ANNOTATION: FROM SEQUENCE TO BIOLOGY

Lincoln Stein

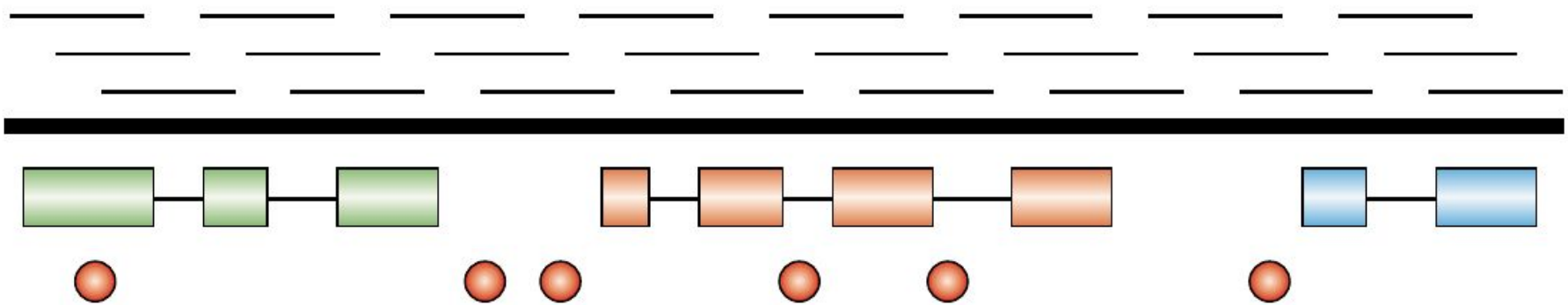
The genome sequence of an organism is an information resource unlike any that biologists have previously had access to. But the value of the genome is only as good as its annotation. It is the annotation that bridges the gap from the sequence to the biology of the organism. The aim of high-quality annotation is to identify the key features of the genome — in particular, the genes and their products. The tools and resources for annotation are developing rapidly, and the scientific community is becoming increasingly reliant on this information for all aspects of biological research.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

What is genome annotation?

Where?

Nucleotide-level annotation



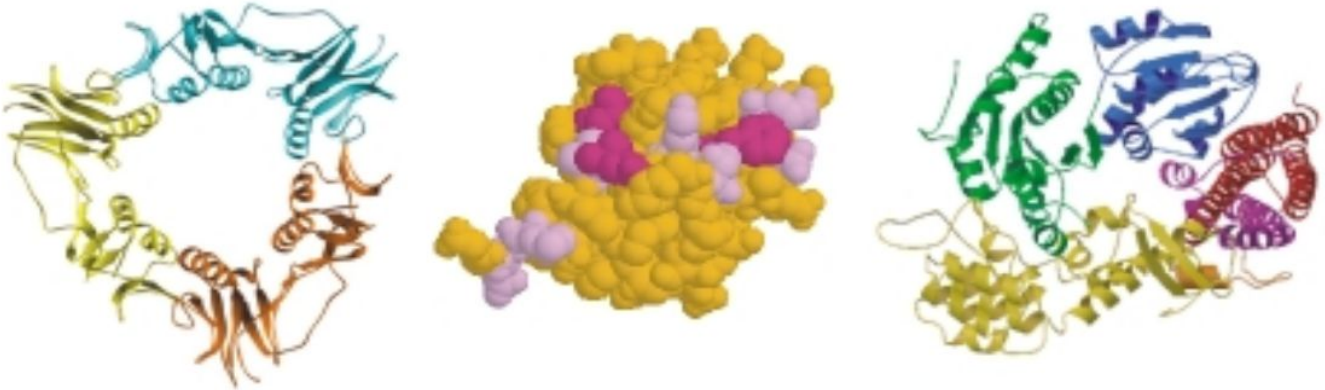
Structura annotation refers to the process of identifying and annotating various structural features within a genome, such as genes, regulatory elements, and other functional elements (gene finding, promoter prediction, etc).

Adapted from Stein L. 2001.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

What is genome annotation?

What?
Protein-level annotation



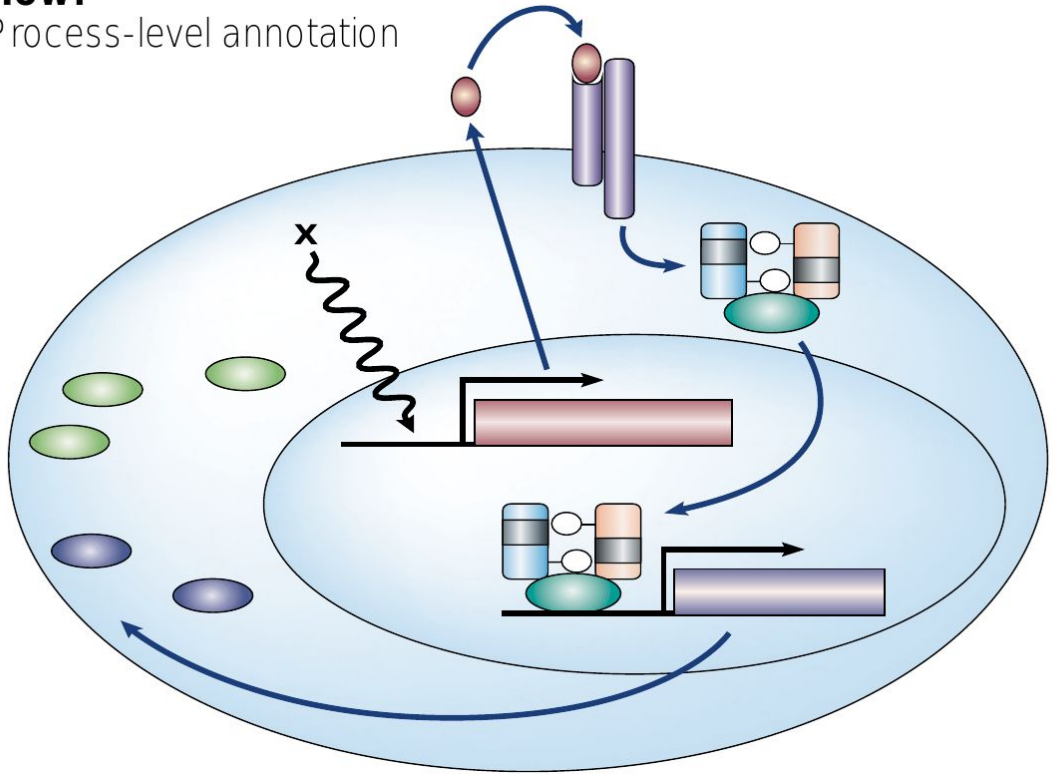
Functional annotation refers to the process of assigning functional information to gene sequences or genomic regions. This information helps researchers understand the biological roles and significance of these sequences (homology-based, domains prediction).

Adapted from Stein L. 2001.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

What is genome annotation?

How?
Process-level annotation

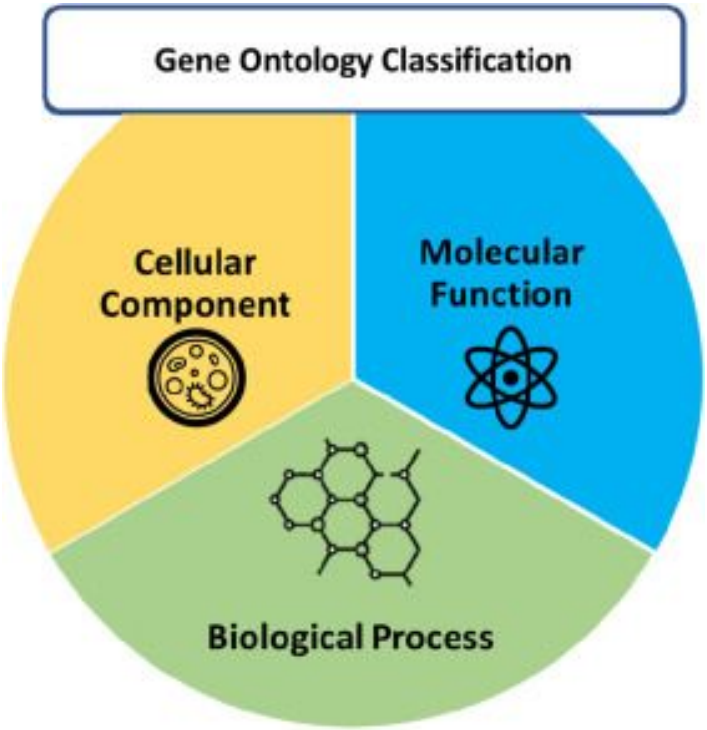


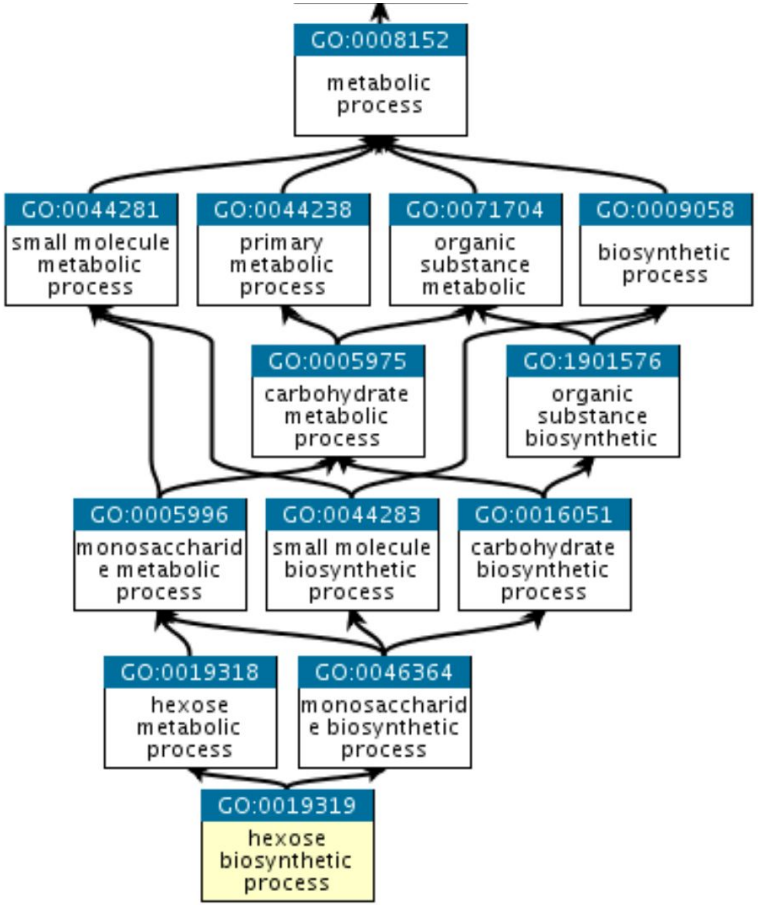
Process level annotation refers to the annotation of genes or genomic regions based on their involvement in specific biological processes. This type of annotation provides information about the functions of genes and their roles in cellular processes, pathways, and systems.

Adapted from Stein L. 2001.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

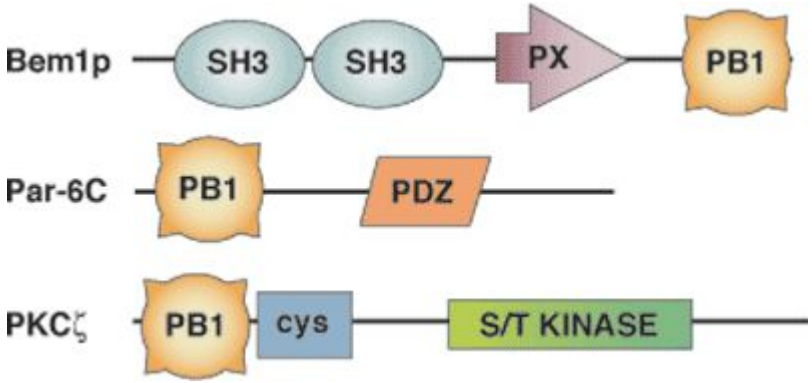
Nombrando entidades y relaciones → ontologías





Buscan estandarizar el uso del lenguaje de las anotaciones (interoperabilidad)

Nombrando entidades y relaciones → ontologías



Molecular function: Molecular-level activities performed by gene products. Molecular function terms describe activities that occur at the molecular level, such as “catalysis” or “transport”.

Cellular Component: The locations relative to cellular structures in which a gene product performs a function, either cellular compartments (e.g., mitochondrion), or stable macromolecular complexes…

Biological process: The larger processes, or ‘biological programs’ accomplished by multiple molecular activities. Examples of broad biological process terms are DNA repair or signal transduction. Examples of more specific terms are pyrimidine nucleobase biosynthetic process or glucose transmembrane transport.

El problema parecería estar resuelto en teoría, pero en la práctica…

BJ www.biochemj.org

BJ Metabolism

Biochem. J. (2010) **425**, 1–11 (Printed in Great Britain) doi:10.1042/BJ20091328

REVIEW ARTICLE

‘Unknown’ proteins and ‘orphan’ enzymes: the missing half of the engineering parts list – and how to find it

Andrew D. HANSON*[1], Anne PRIBAT*, Jeffrey C. WALLER* and Valérie DE CRÉCY-LAGARD†

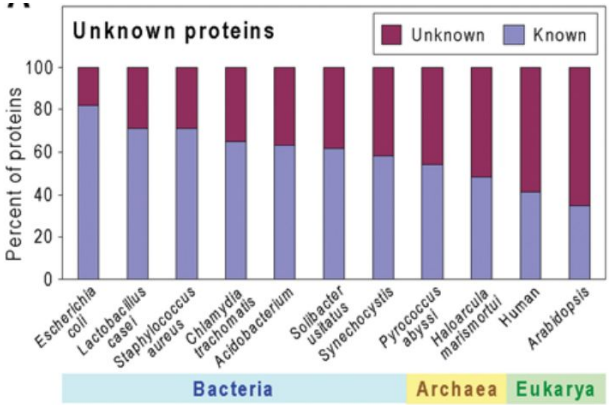
*Horticultural Sciences Department, University of Florida, Gainesville, FL 32611, U.S.A., and †Microbiology and Cell Science Department, University of Florida, Gainesville, FL 32611, U.S.A.

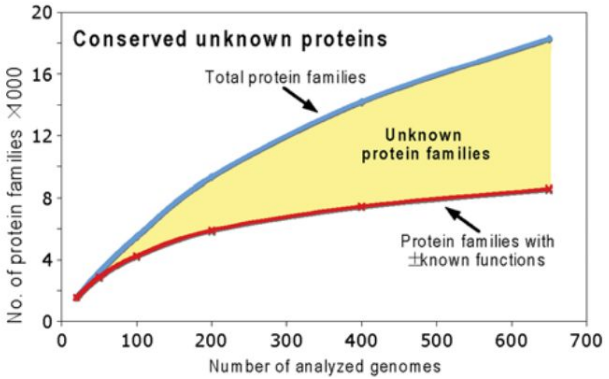
'Unknown' proteins and orphan enzymes

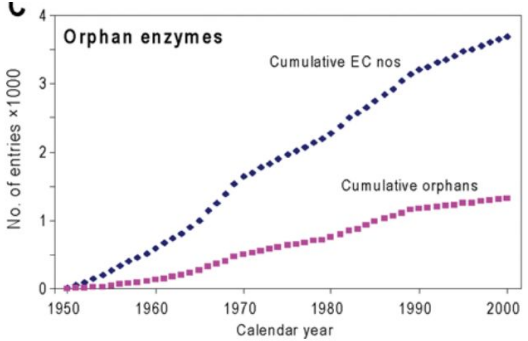
The elephant in the room

Two classes:

- **Hypothetical protein / putative conserved:** Sequences that "looks" like genes, are conserved across liange but nobody knows their function (in the most brad sense).
- **Orphan functions:** known functions (e.g. enzymatic activity) that nobody knows a candidate gene encoding that function.







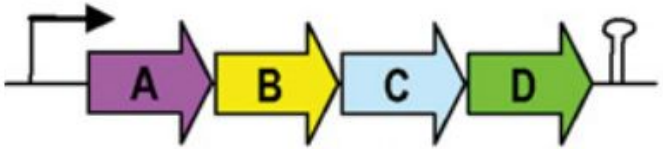
Adapted from Hanson et 2010.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

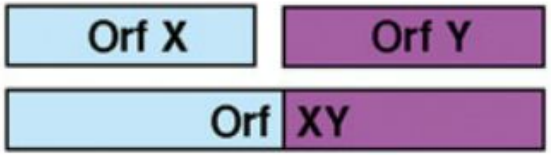
The predictive power of comparative genomics

The 'guilt by association' principle

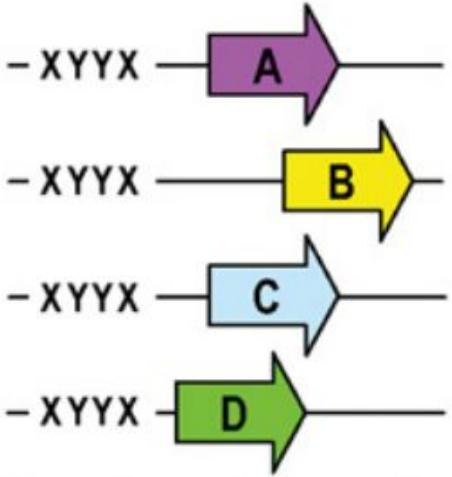
Types of associations in comparative genomics: **genomic evidence**



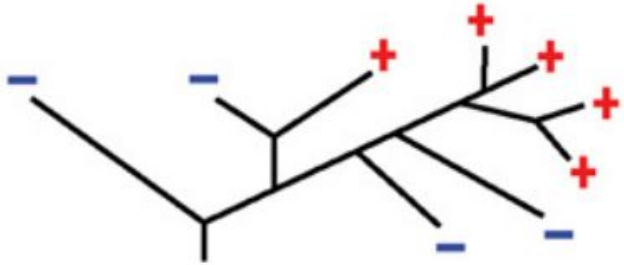
Gene clustering



Gene fusion



Shared regulatory sites



Phylogenetic occurrence

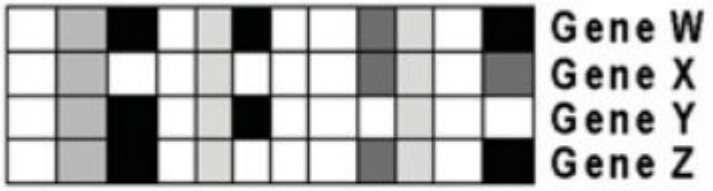
Adapted from Hanson et 2010.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

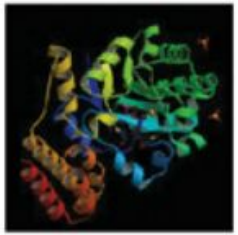
The predictive power of comparative genomics

The 'guilt by association' principle

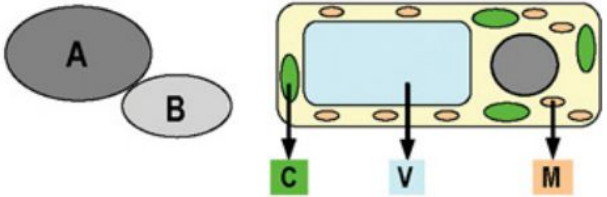
Associations based on post-genomic resources: **post-genomic evidence**



Co-expression

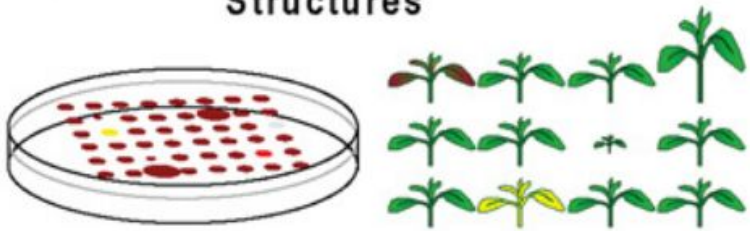


Structures



Protein-protein interactions

Organelle proteomes

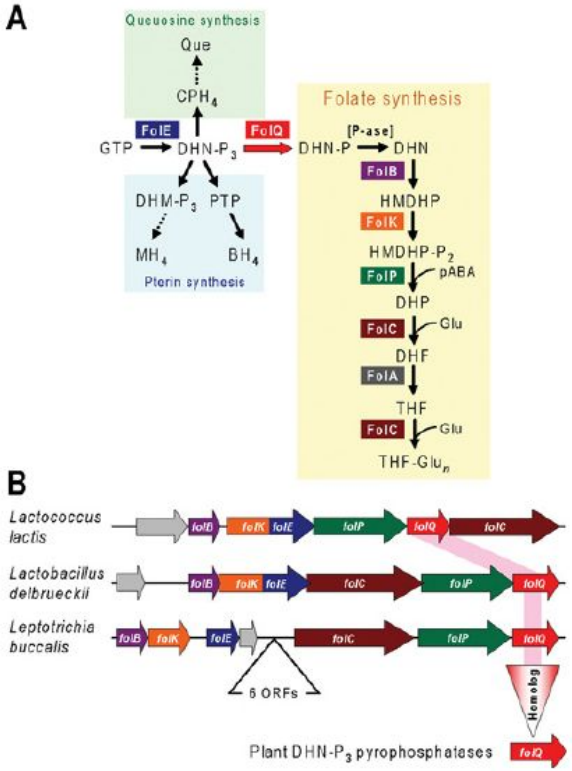


Essentiality & other phenome data

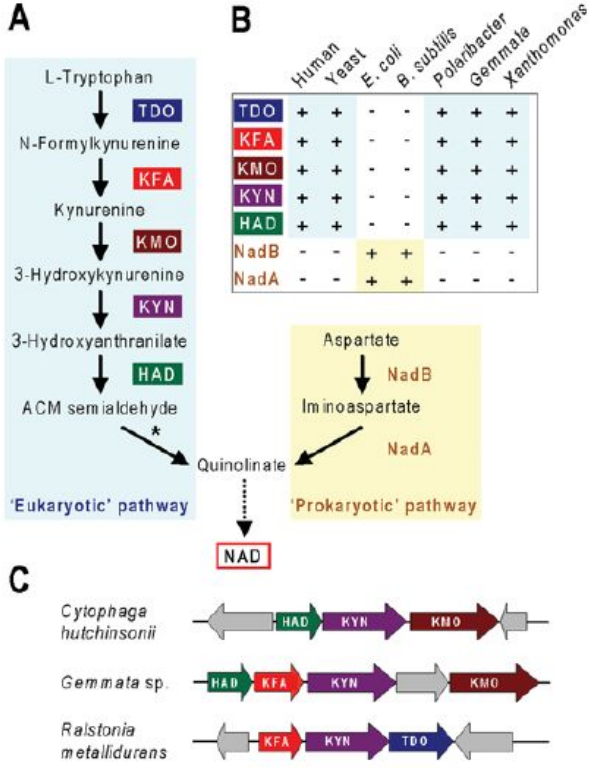
Adapted from Hanson et 2010.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Examples



A missing folate biosynthesis enzyme



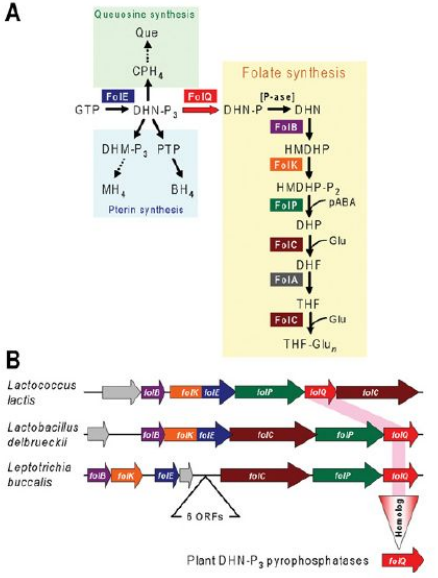
Tryptophan to quinolinate in NAD biosynthetic pathway

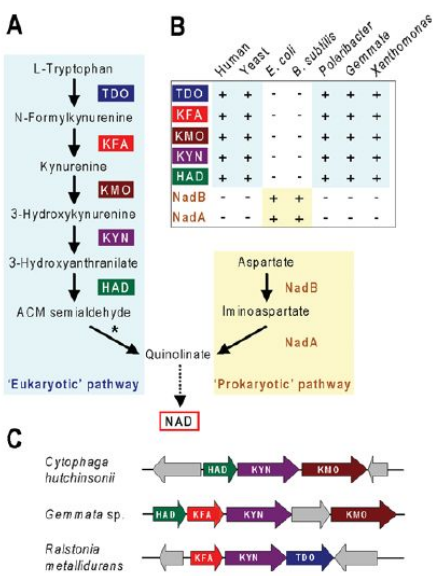
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

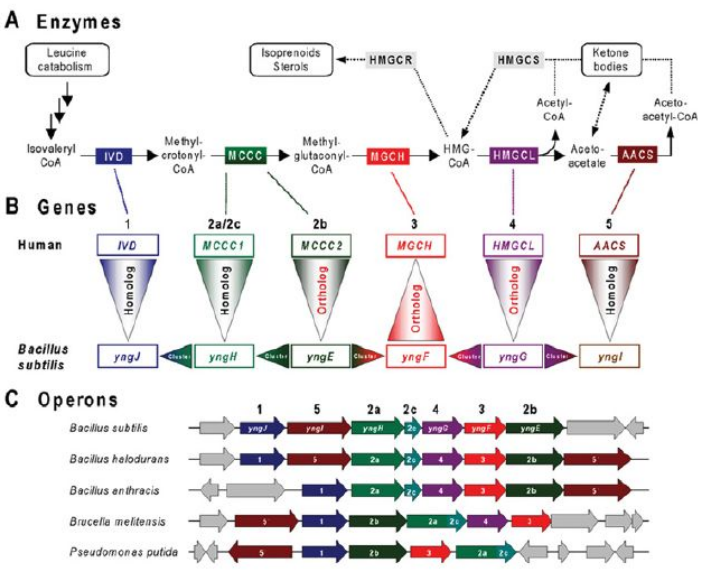
The predictive power of comparative genomics: the 'guilt by association' principle

Example 1: a missing folate biosynthesis enzyme

Example 2: the tryptophan to quinolinate route in NAD synthesis







Machine Learning Based Approaches

A. L. Samuel

Some Studies in Machine Learning Using the Game of Checkers

1959

Abstract: Two machine-learning procedures have been investigated in some detail using the game of checkers. Enough work has been done to verify the fact that a computer can be programmed so that it will learn to play a better game of checkers than can be played by the person who wrote the program. Furthermore, it can learn to do this in a remarkably short period of time (8 or 10 hours of machine-playing time) when given only the rules of the game, a sense of direction, and a redundant and incomplete list of parameters which are thought to have something to do with the game, but whose correct signs and relative weights are unknown and unspecified. The principles of machine learning verified by these experiments are, of course, applicable to many other situations.

Volume 93, 1992

Previous Article | Next Article

From the journal:
Faraday Discussions

Modelling the structure and function of enzymes by machine learning

Michael J. E. Sternberg, Richard A. Lewis, Ross D. King and Stephen Muggleton

Abstract

A machine learning program, GOLEM, has been applied to two problems: (1) the prediction of protein secondary structure from sequence and (2) modelling a quantitative structure-

PubMed.gov

machine learning

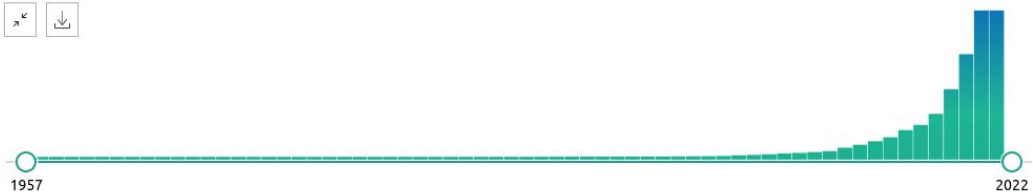
Search

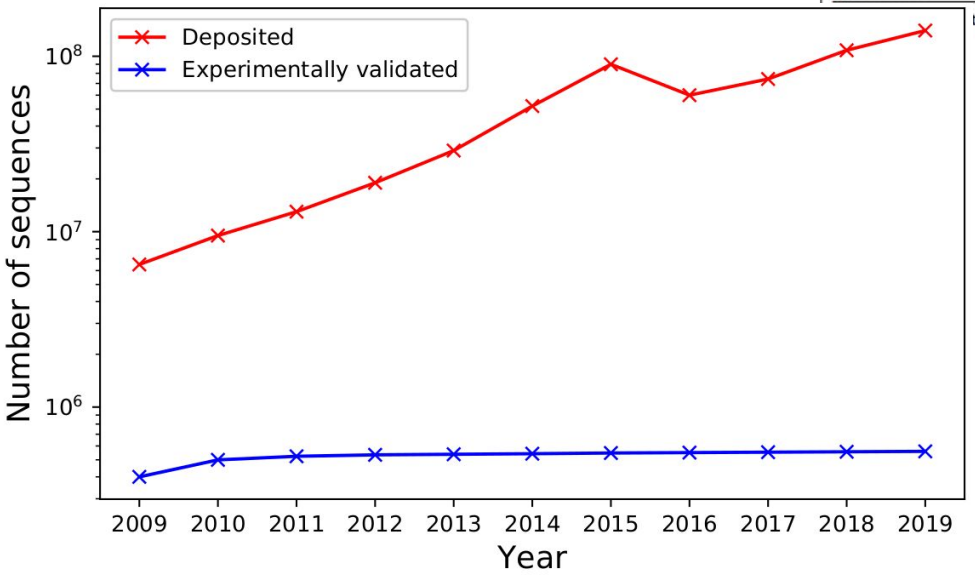
Advanced Create alert Create RSS User Guide

Save Email Send to Sorted by: Most recent Display options

RESULTS BY YEAR

71,453 results





Article

Unraveling the functional dark matter through global metagenomics

https://doi.org/10.1038/s41586-023-06583-7
Received: 18 March 2022
Accepted: 30 August 2023
Published online: 11 October 2023
Open access

Georgios A. Pavlopoulos[1,2,3✉], Fotis A. Baltoumas[1], Sirui Liu[4], Oguz Selvitopi[5], Antonio Pedro Camargo[2], Stephen Nayfach[2], Ariful Azad[6], Simon Roux[2], Lee Call[2], Natalia N. Ivanova[2], I. Min Chen[2], David Paez-Espino[2], Evangelos Karatzas[1], Novel Metagenome Protein Families Consortium*, Ioannis Iliopoulos[7], Konstantinos Konstantinidis[8], James M. Tiedje[9], Jennifer Pett-Ridge[10], David Baker[11,12,13], Axel Visel[2], Christos A. Ouzounis[2,14,15], Sergey Ovchinnikov[4], Aydin Buluç[5,16] & Nikos C. Kyrpides[2✉]

Metagenomes encode an enormous diversity of proteins, reflecting a multiplicity of functions and activities[1,2]. Exploration of this vast sequence space has been limited to a comparative analysis against reference microbial genomes and protein families derived from those genomes. Here, to examine the scale of yet untapped functional diversity beyond what is currently possible through the lens of reference genomes, we develop a computational approach to generate reference-free protein families from the sequence space in metagenomes. We analyse 26,931 metagenomes and identify 1.17 billion protein sequences longer than 35 amino acids with no similarity to any sequences from 102,491 reference genomes or the Pfam database[3]. Using massively parallel graph-based clustering, we group these proteins into 106,198 novel sequence clusters with more than 100 members, doubling the number of protein families obtained from the reference genomes clustered using the same approach. We

Bonetta & Valentino, 2019

General workflow

Genes with known functions

Training & evaluation

Sequence features
Structural features
Expression patterns
Interaction networks
Phylogenies
Gene location

Feature engineering, feature selection, embeddings

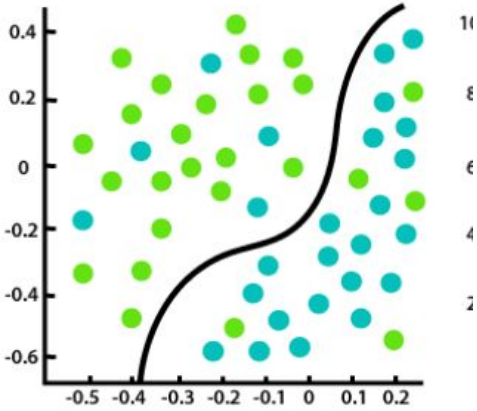
ML model

Algorithms & architectures

Functional predictions

List of candidate genes

Experimental validation

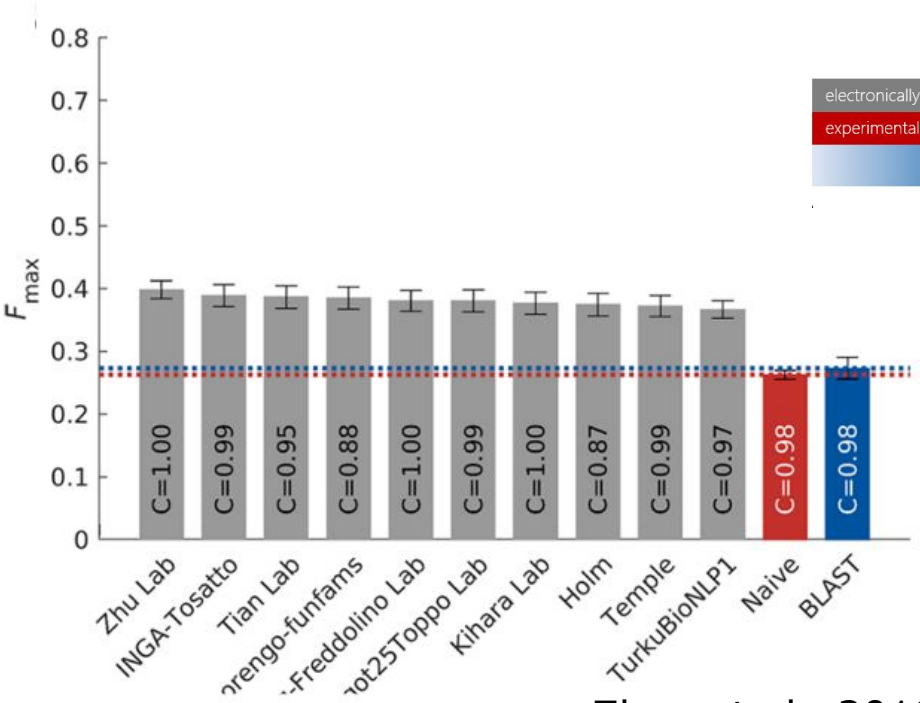


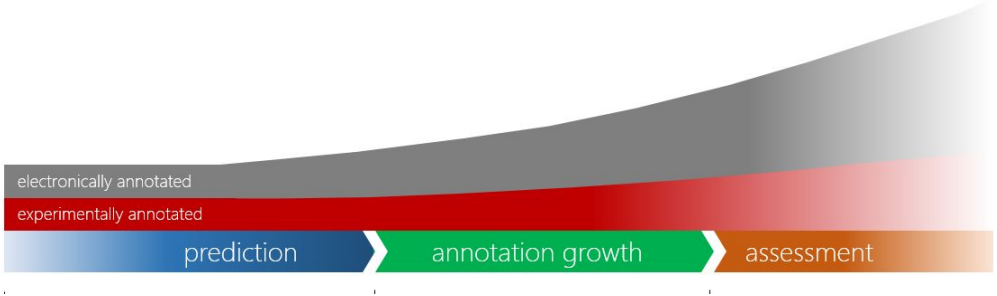
Some characteristics of gene function prediction as a classification problem

- Multiclass classification: large number of functional classes.
- Multilabel classification: multiple annotations for each gene.
- Labels at different level of reliability.
- Hierarchical relationships between classes (structured output).
- Unbalanced classification.
- Different strategies to choose negative examples.
- Multiple sources of relevant data available.
- Data are usually complex and noisy.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

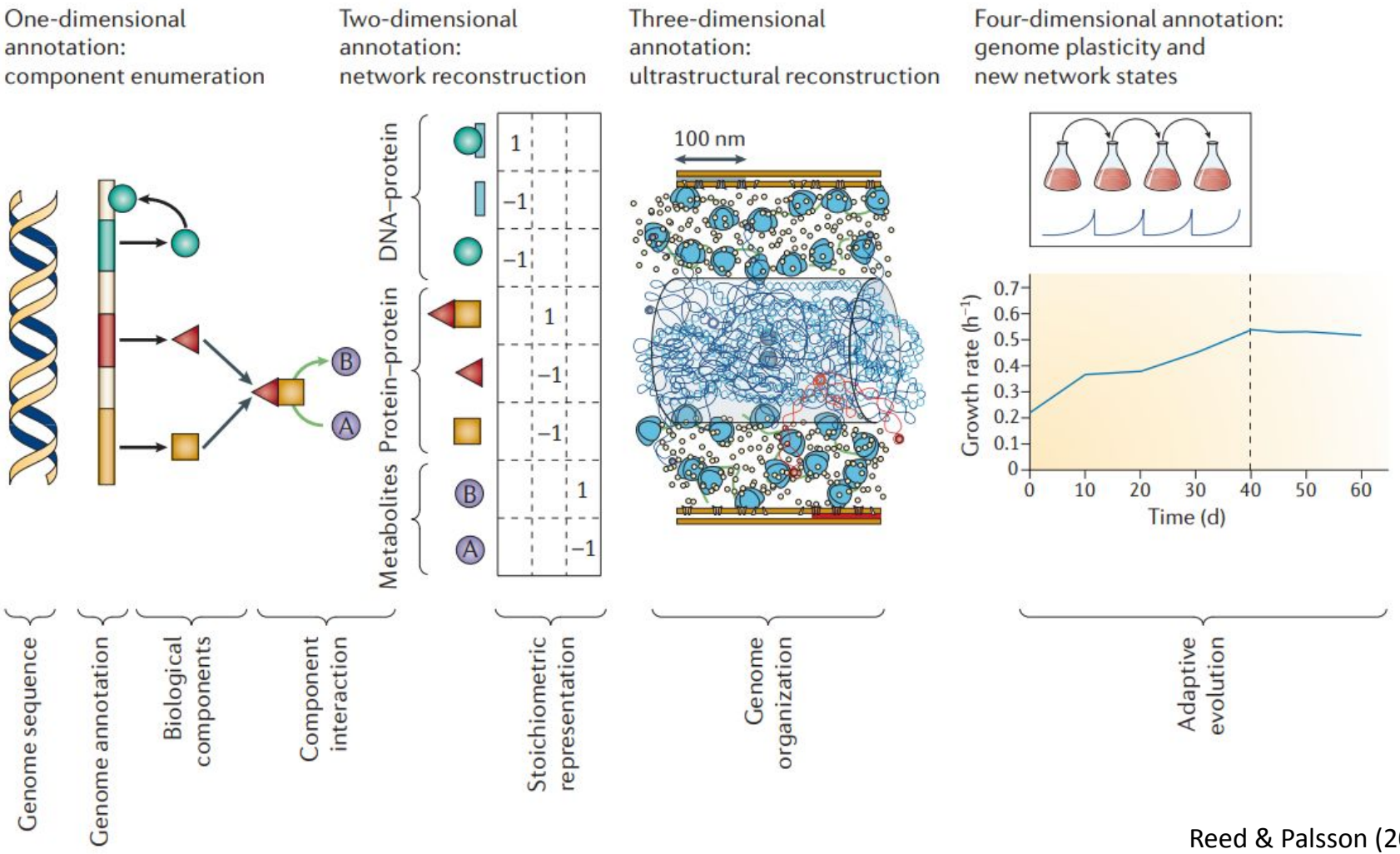
An open problem



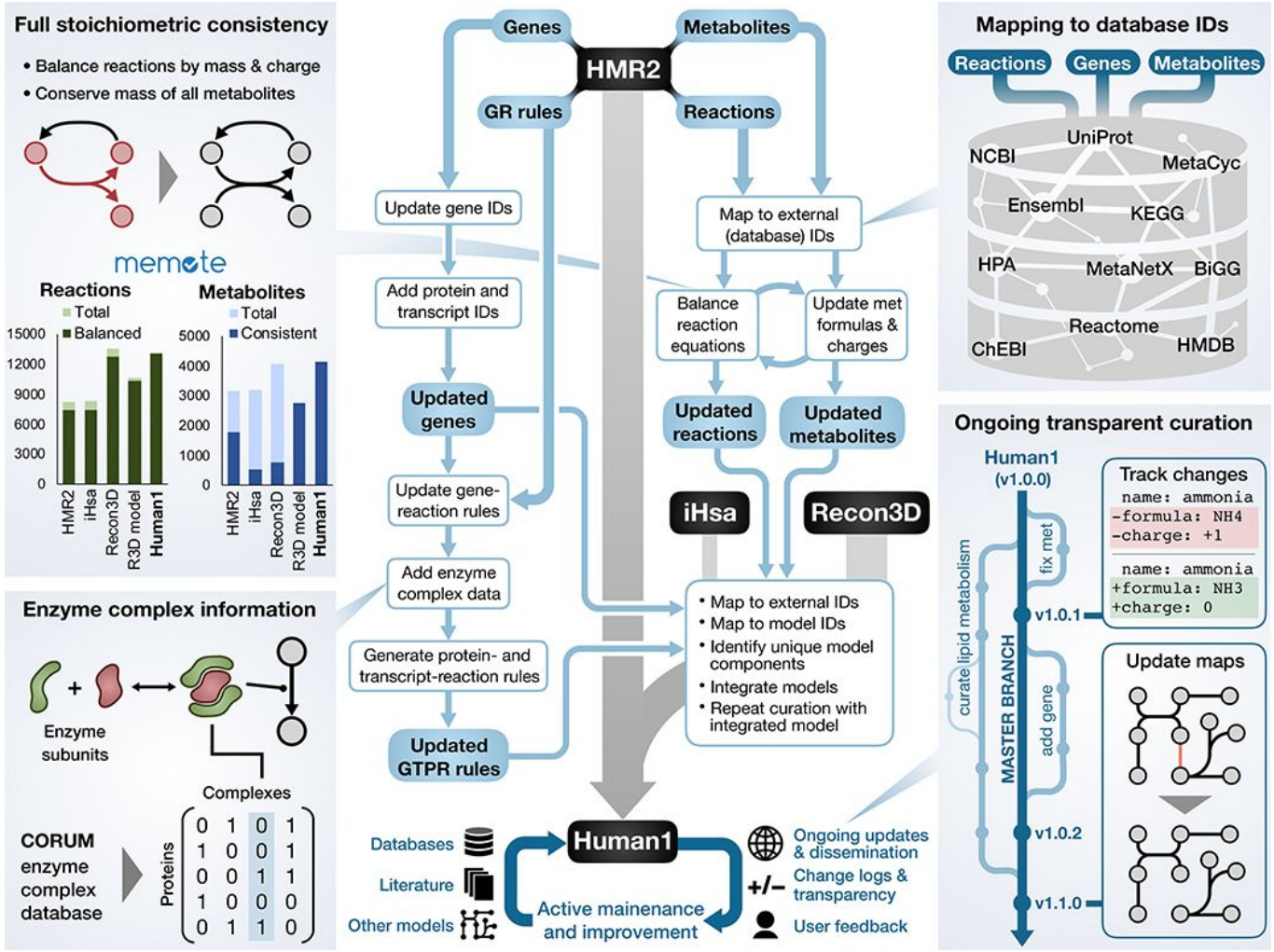


Zhou et al., 2019

Reconstructing Molecular networks



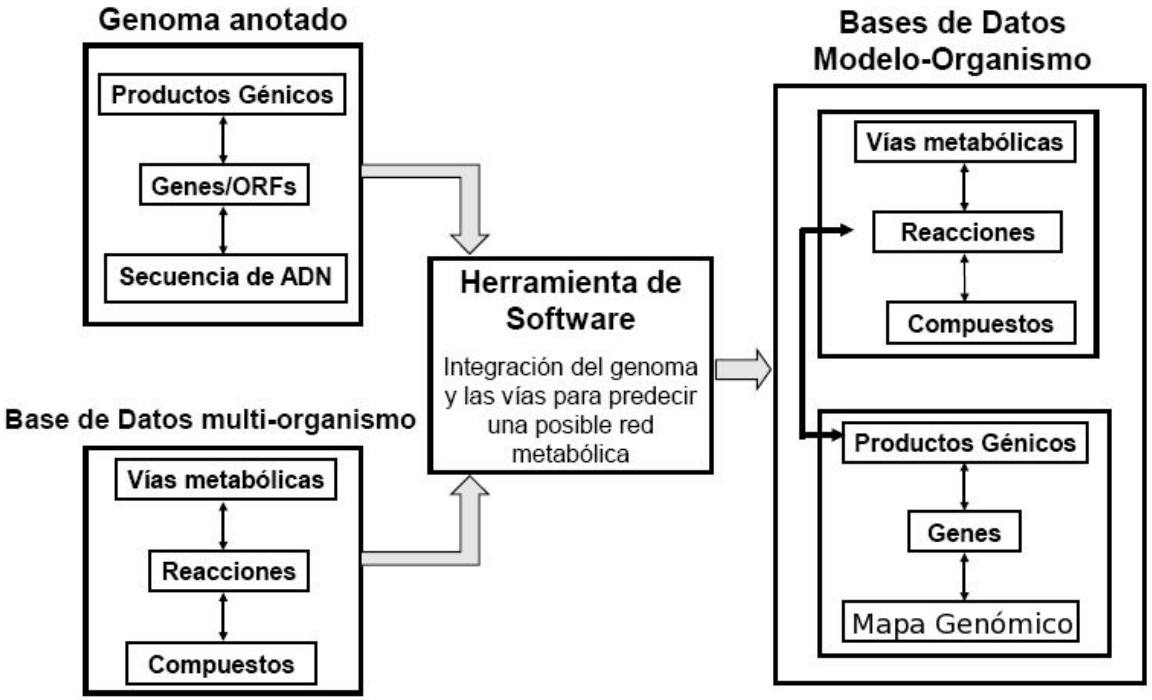
Reed & Palsson (2006)

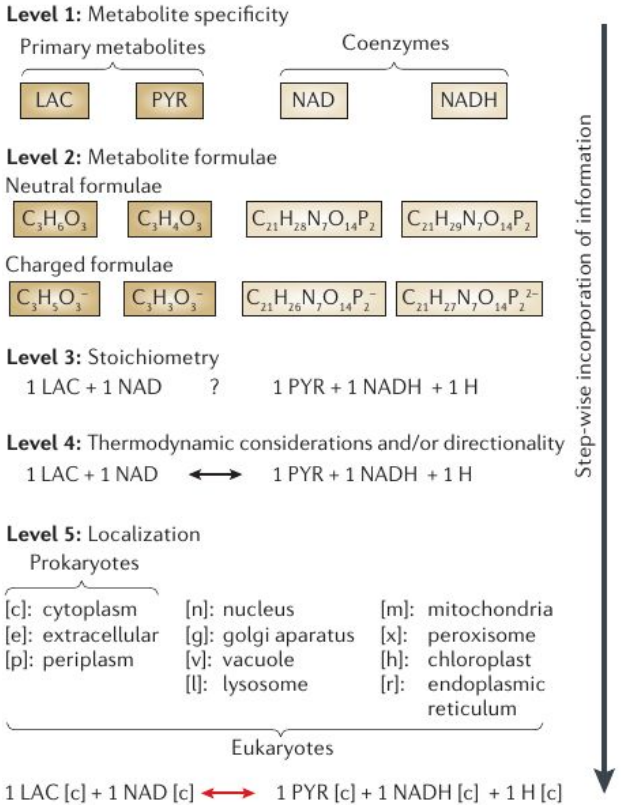


Robinson et al. (2021)

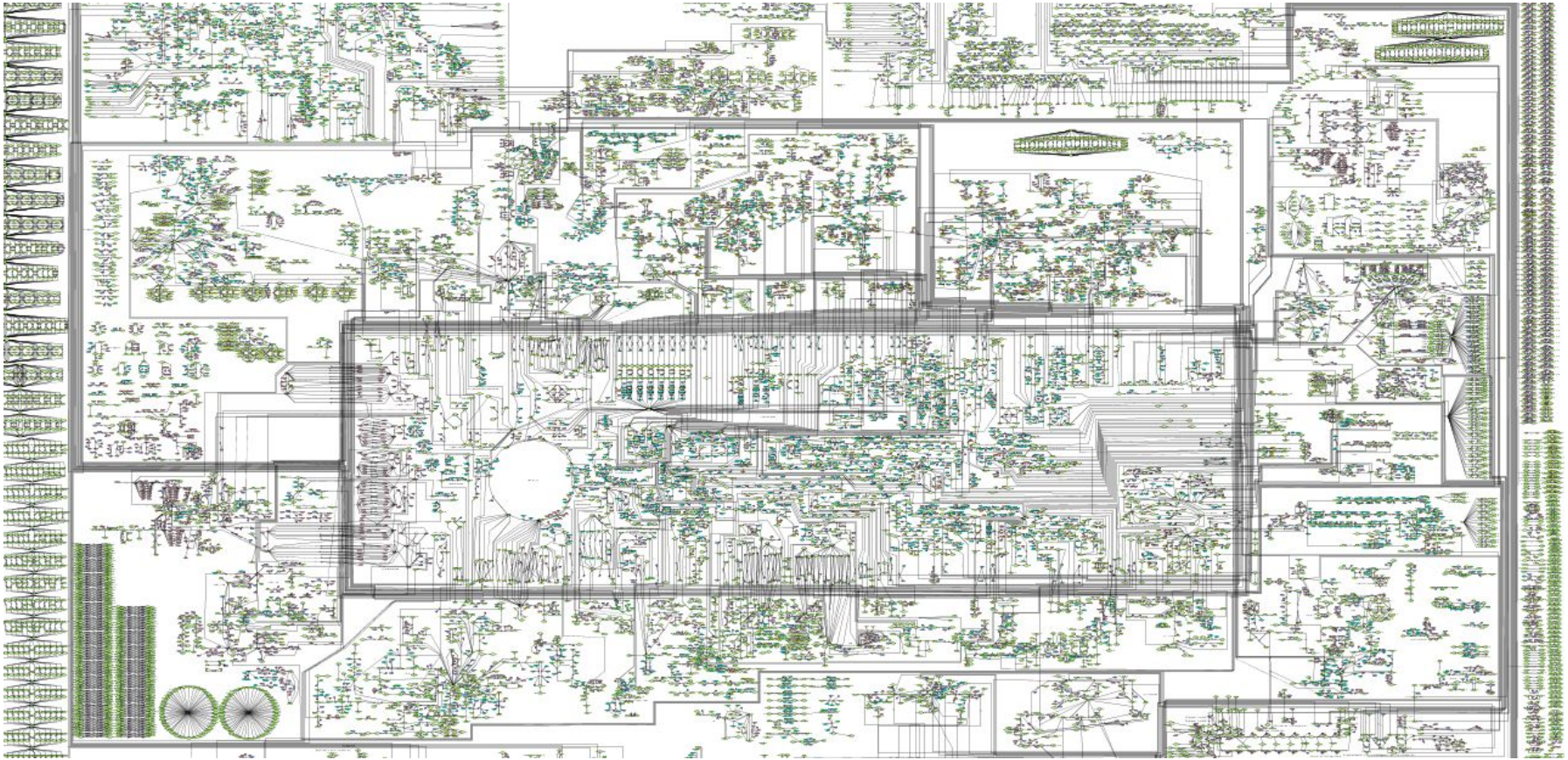
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Reconstrucción metabólica





Virtual Human Metabolic:



https://www.vmh.life/minerva/

Genome-Scale Model of Human Metabolism

Table 1

Statistics of currently published generic human GEMs.

Generic GEMs	Genes	Metabolites[a]	Reactions[a]	Features
RECON1	1496	1509	3744	Manually reconstructed from bibliomics data
EHMN	2322	2671	2823	Manually reconstructed from bibliomics data
RECON2	1789	2626	7440	Merging EHMN and HepatoNet1 with RECON1
RECON 2.2	1675	5324	7785	Reconstructed by integrating previous versions, with emphasis on mass and charge balance
HMR1.0	1512	3397	4144	Reconstructed based on RECON1, EHMN, HumanCyc and KEGG
HMR2.0	3765	3160	8181	Reconstructed based on HMR1, with additional emphasis on lipid metabolism by integrating iAdipocytes1809, KEGG, Lipidomics Gateway
Recon3D	2248	5835	10600	Reconstructed based on RECON2 and includes mapping to 3D structure of proteins through PDB ids
Human1	3625	10138 (4164)	13417	Integrated and extensively curated the most recent human metabolic models to construct a consensus GEM, Human1

- Several options available (all derived from RECON1)
- Human1 is most recent version

Modified from Swainston, N., et al (2016). Metabolomics, 12(7), 109.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Discusión

¿Qué entendemos por “anotar”?

Función molecular; contexto y contribución al fenotipo; historia evolutiva..

¿Se reduce a enumerar y describir componentes y sus interacciones?

¿Qué pasa con las interacciones?

¿Cuando consideramos que una entidad o proceso está anotada?

Función molecular; contexto y contribución al fenotipo; historia evolutiva..

¿Se podrá completar el catálogo algún día?

¿Tiene sentido la pregunta anterior?

¿Open ended evolution?

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Referencias

Anotación de genomas y reconstrucción de redes

- Genome annotation from sequence to biology (Stein - 2001)
- Unknown proteins and orphan enzymes the missing half of the engineering parts list – and how to find it (Hanson et al - 2010)
- The model organism as a system: integrating 'omics' data sets (Joyce & Palsson, 2006)
- Toward multidimensional genome annotation (Reed et. al, 2006)

Perspectivas desde el pasado reciente

- From molecular to modular cell biology (Hartwell et al - 1999)
- What lies beyond bioinformatics? (Palsson - 1997)