



Bioinformatics Master's Course

Genome Analysis (Integrative Bioinformatics & Genomics)

Lecture 1: Introduction

Centre for Integrative Bioinformatics VU (IBIVU)
Faculty of Exact Sciences / Faculty of Earth and Life Sciences
<http://ibi.vu.nl>, heringa@few.vu.nl, 87649 (Heringa), Room P1.28

Other teachers (assistants) in the course

- Anton Feenstra, UD (1/09/05)

- Bart van Houte – PhD (1/09/04)

- Walter Pirovano – PhD (1/09/05)

- Thomas Binsl - PhD (18/6/06)

- Bernd Brandt (18/06/07)



Issues in data analysis

- Pattern recognition
 - Supervised/unsupervised learning
 - Types of data, data normalisation, lacking data
 - Search image
 - Similarity/distance measures
 - Clustering
 - Principal component analysis

Protein Science (the ‘doers’ in the cell)

- Protein
 - Folding
 - Structure and function
 - Protein structure prediction
 - Secondary structure
 - Tertiary structure
 - Function
 - Post-translational modification
 - Prot.-Prot. Interaction -- Docking algorithm
 - Molecular dynamics/Monte Carlo

Central Bioinformatics issue: Sequence Analysis

- Sequence analysis
 - Pairwise alignment
 - Dynamic programming (NW, SW, shortcuts)
 - Multiple alignment
 - Combining information
 - Database/homology searching (Fasta, Blast, Statistical issues-E/P values)

Bioinformatics algorithms for Genomics

- Gene structure and gene finding algorithms
- Algorithms to integrate Genomics databases:
 - Sequencing projects
 - Expression data, Nucleus to ribosome, translation, etc.
 - Proteomics, Metabolomics, Physiomics
 - Databases
 - DNA, EST
 - Protein sequence (SwissProt)
 - Protein structure (PDB)
 - Microarray data
 - Proteomics
 - Mass spectrometry/NMR/X-ray

Gathering knowledge

- Anatomy, architecture



Rembrandt,
1632

- Dynamics, mechanics



Newton,
1726

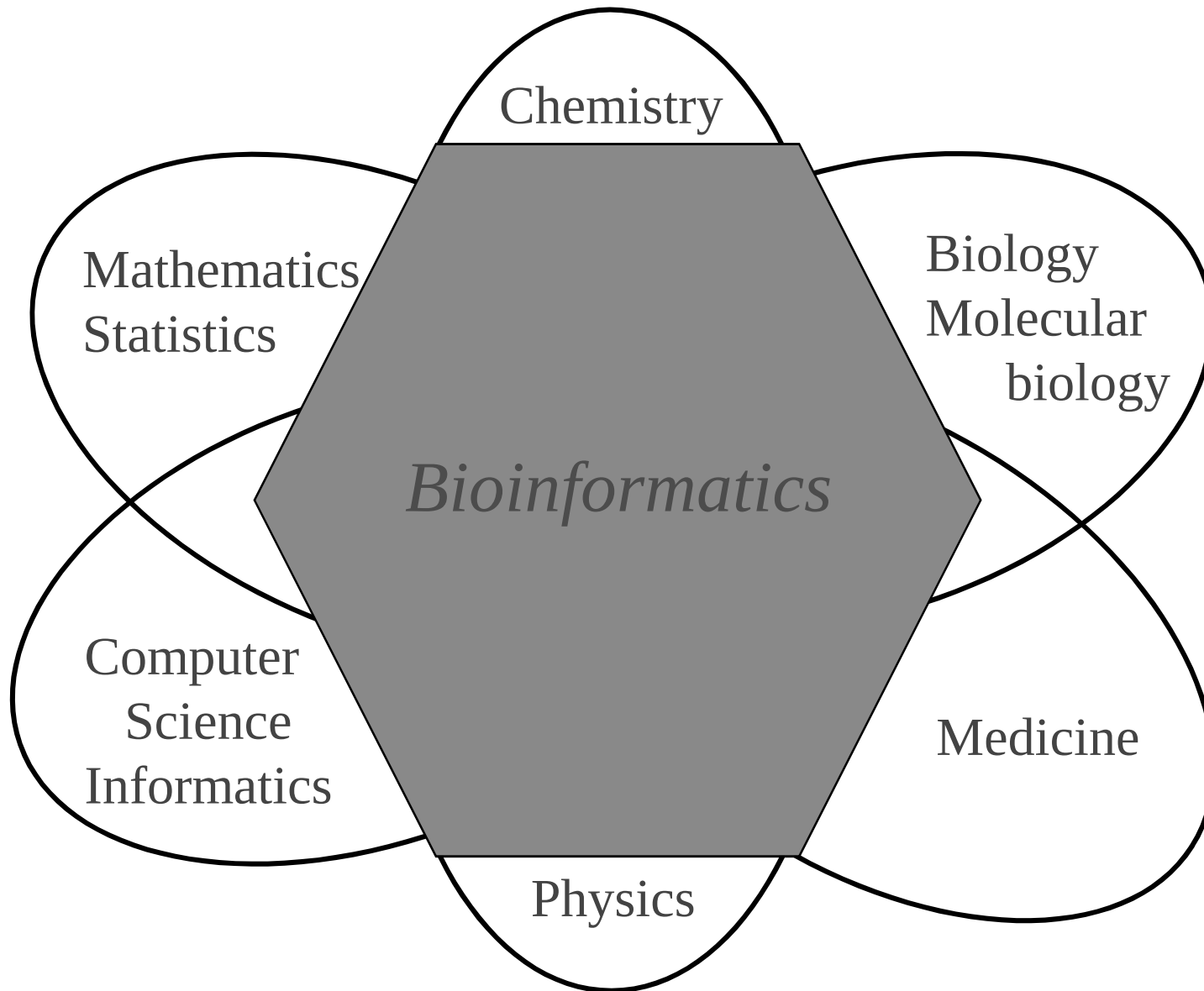
- Informatics

(Cybernetics – Wiener, 1948)

(Cybernetics has been defined as the science of control in machines and animals, and hence it applies to technological, animal and environmental systems)

- Genomics, bioinformatics

Bioinformatics



Bioinformatics

“Studying informational processes in biological systems”
(Hogeweg, early 1970s)

- No computers necessary
- Back of envelope OK



*“Information technology
applied to the management and
analysis of biological data”
(Attwood and Parry-Smith)*

Applying algorithms with mathematical formalisms in
biology (genomics)

*Not good: biology and biological knowledge is crucial for making
meaningful analysis methods!*

Bioinformatics in the olden days

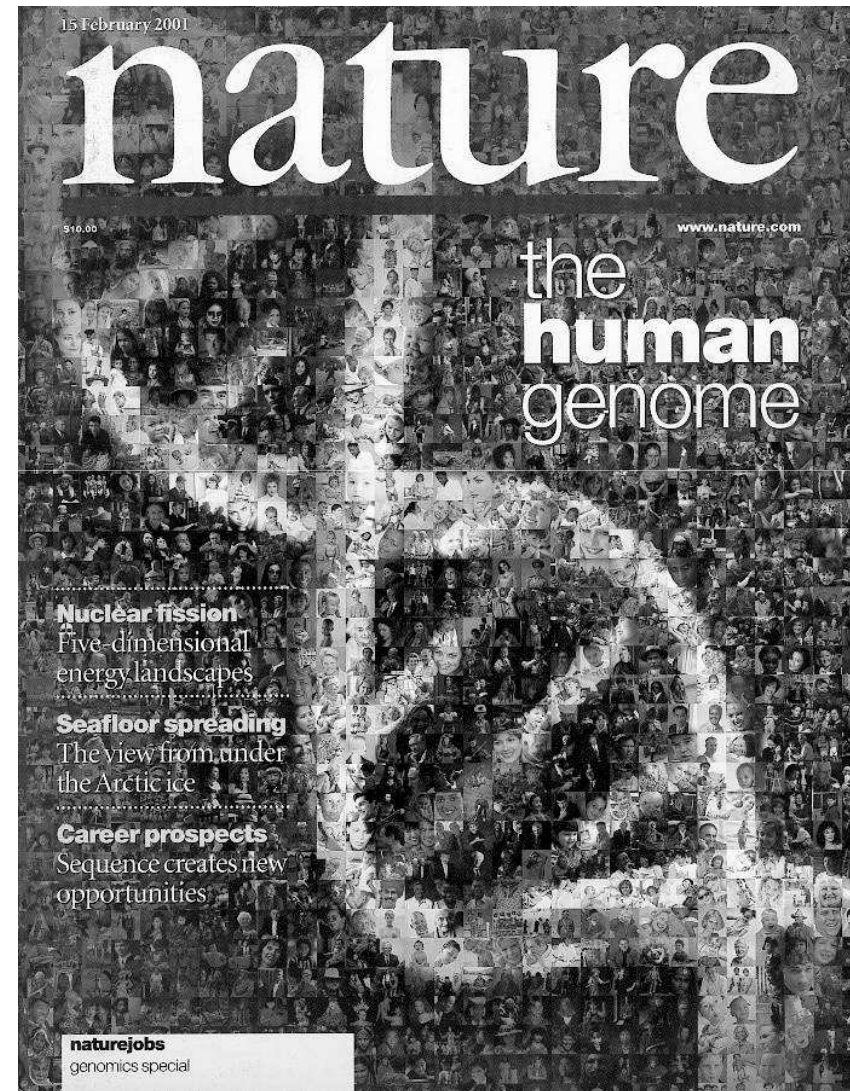
- Close to Molecular Biology:
 - (Statistical) analysis of protein and nucleotide structure
 - Protein folding problem
 - Protein-protein and protein-nucleotide interaction
- Many essential methods were created early on (BG era)
 - Protein sequence analysis (pairwise and multiple alignment)
 - Protein structure prediction (secondary, tertiary structure)

Bioinformatics in the olden days (Cont.)

- Evolution was studied and methods created
 - Phylogenetic reconstruction (clustering – e.g., Neighbour Joining (NJ) method)

But then the big bang....

The Human Genome -- 26 June 2000



The Human Genome -- 26 June 2000

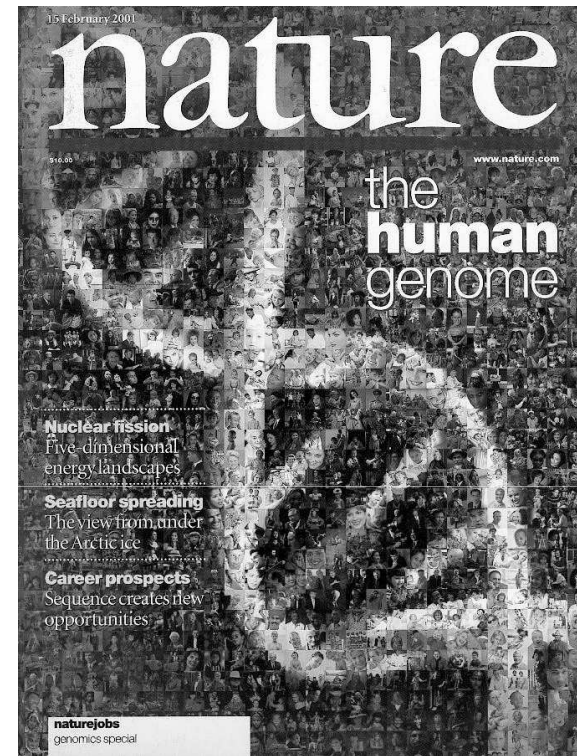
“Without a doubt, this is the most important, most wondrous map ever produced by humankind.”

U.S. President Bill Clinton on 26 June 2000 during a press conference at the White House.

The Human Genome -- 26 June 2000



Dr. Craig Venter
Celera Genomics
-- Shotgun method



Francis Collins (USA)/
Sir John Sulston (UK)
Human Genome Project

Human DNA

- There are at least 3bn (3×10^9) nucleotides in the nucleus of almost all of the trillions (3.2×10^{12}) of cells of a human body (an exception is, for example, red blood cells which have no nucleus and therefore no DNA) – a total of $\sim 10^{22}$ nucleotides!
- Many DNA regions code for proteins, and are called genes (1 gene codes for 1 protein as a base rule, but the reality is a lot more complicated)
- Human DNA contains $\sim 26,000$ expressed genes
- Deoxyribonucleic acid (DNA) comprises 4 different types of nucleotides: adenine (A), thiamine (T), cytosine (C) and guanine (G). These nucleotides are sometimes also called bases

Human DNA (Cont.)

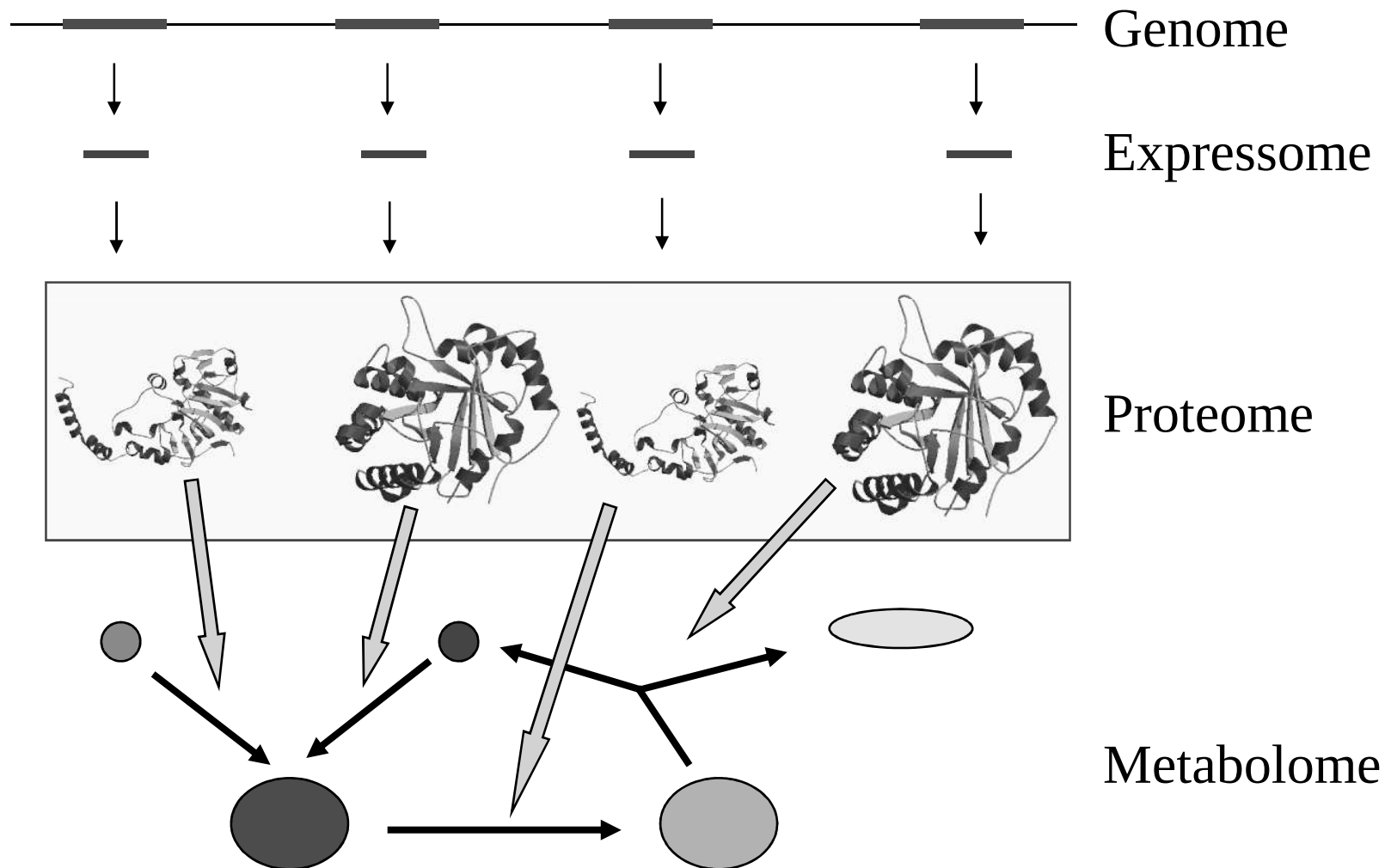
- All people are different, but the DNA of different people only varies for 0.1% or less. Evidence in current genomics studies (Single Nucleotide Polymorphisms or SNPs) imply that on average only 1 letter out of 1400 is different between individuals. Over the whole genome, this means that 2 to 3 million letters would differ between individuals.
- The structure of DNA is the so-called double helix, discovered by Watson and Crick in 1953, where the two helices are cross-linked by A-T and C-G base-pairs (nucleotide pairs – so-called Watson-Crick base pairing).

Modern bioinformatics is closely associated with genomics

- The aim is to solve the genomics information problem
- Ultimately, this should lead to biological understanding how all the parts fit (DNA, RNA, proteins, metabolites) and how they interact (gene regulation, gene expression, protein interaction, metabolic pathways, protein signalling, etc.)
- Genomics will result in the “parts list” of the genome, crucial for cell functioning

Functional Genomics

From gene to networked function



Three new interdisciplinary fields
closely connected to Bioinformatics:

- Translational Medicine
- Systems Biology
- Neurobiology/Neuroinformatics

Translational Medicine

- “From bench to bed side”
- Genomics data to patient data
- Integration

Systems Biology

is the study of the interactions between the components of a biological system, and how these interactions give rise to the function and behaviour of that system (for example, the enzymes and metabolites in a metabolic pathway). The aim is to quantitatively understand the system and to be able to predict the system's time processes

- the interactions are nonlinear
- the interactions give rise to emergent properties, i.e. properties that cannot be explained by the components in the system

Systems Biology

understanding is often achieved through modeling and simulation of the system's components and interactions.

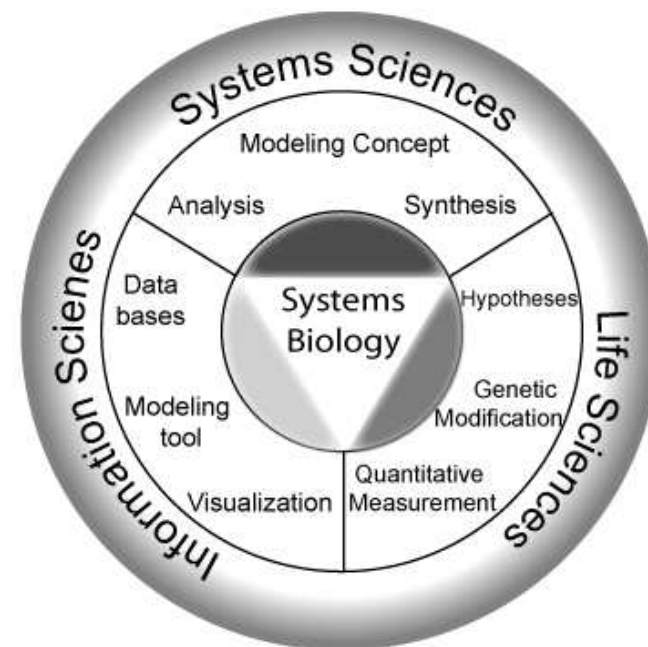
Many times, the 'four Ms' cycle is adopted:

Measuring

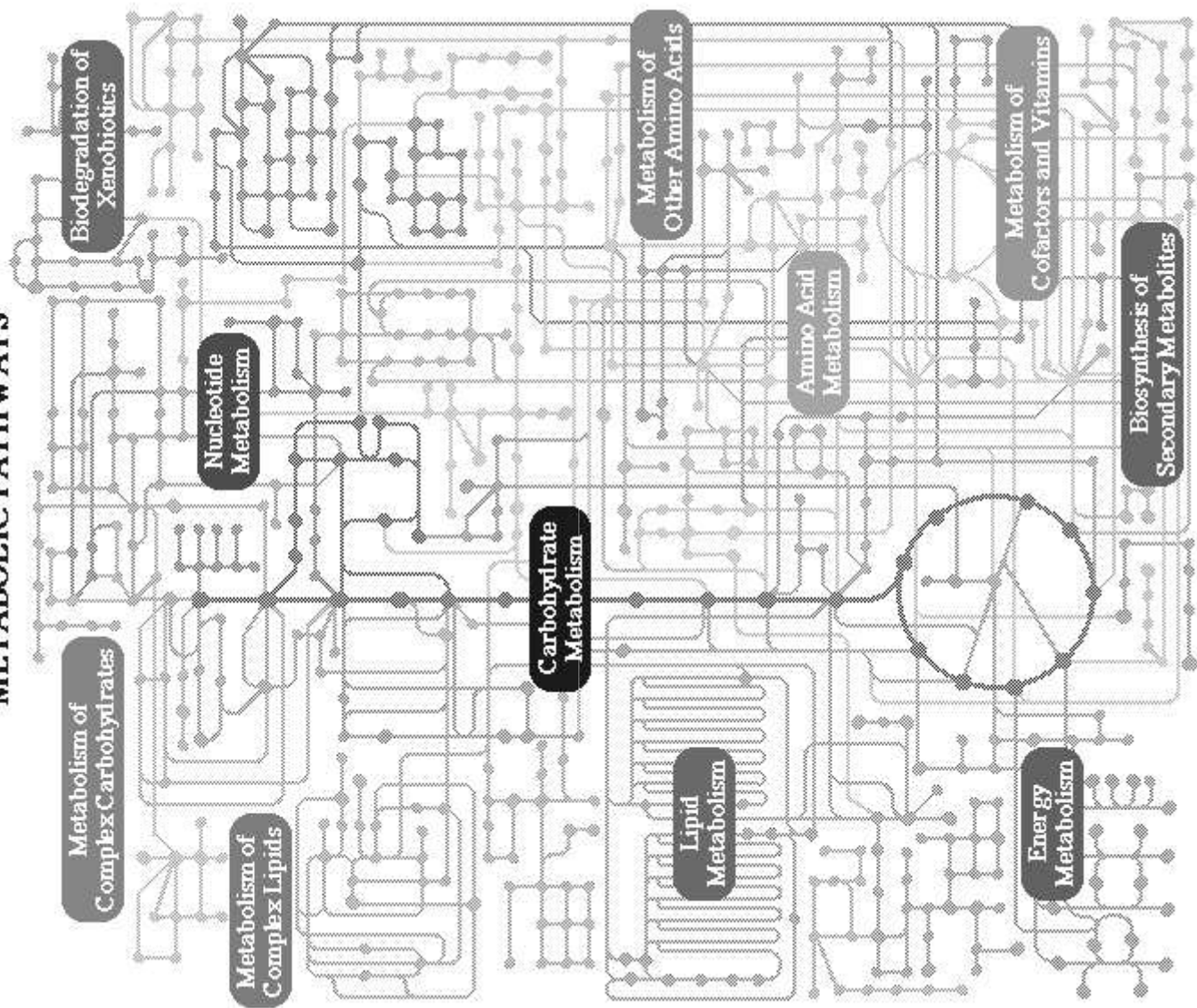
Mining

Modeling

Manipulating

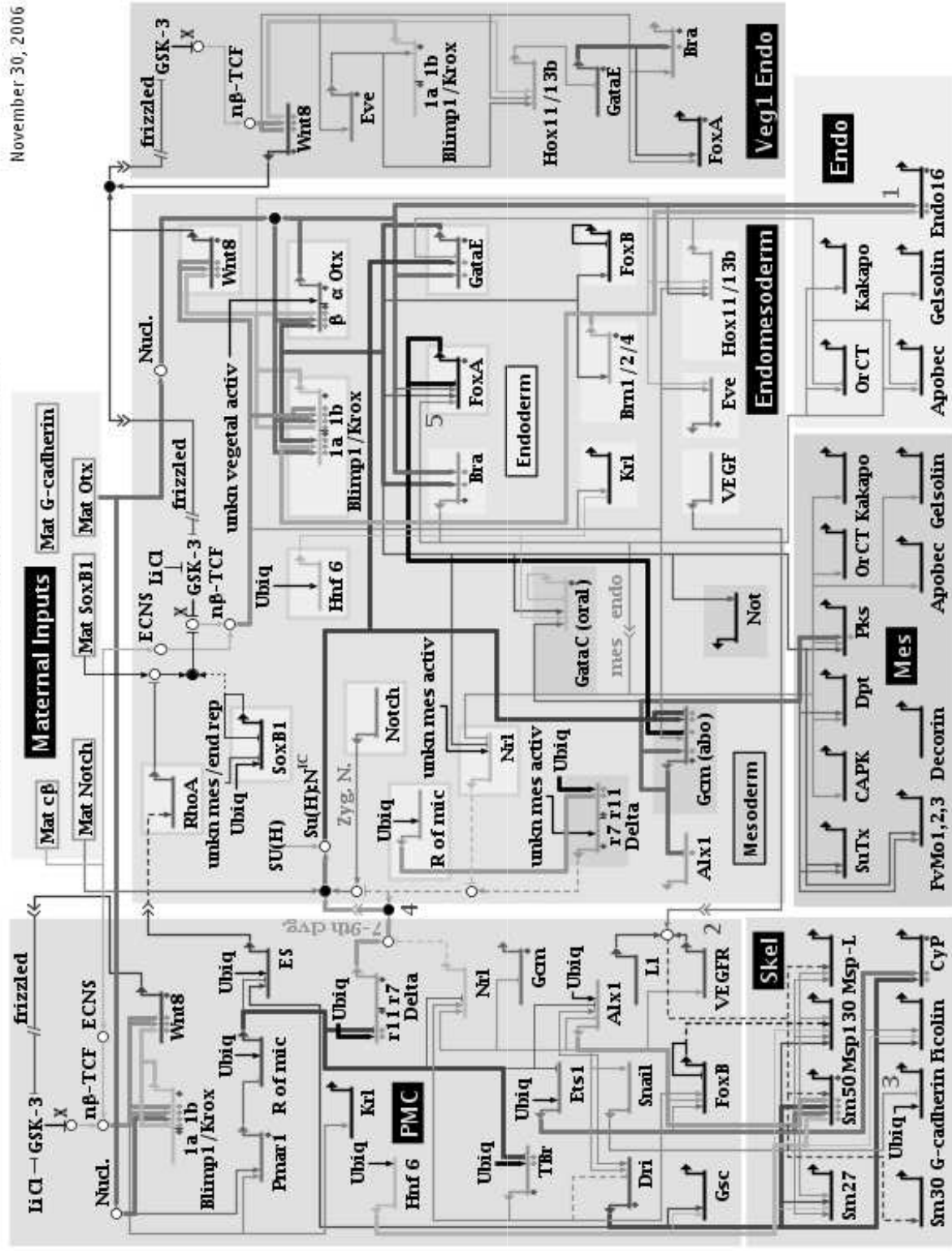


METABOLIC PATHWAYS



Endomesoderm Specification to 30 Hours

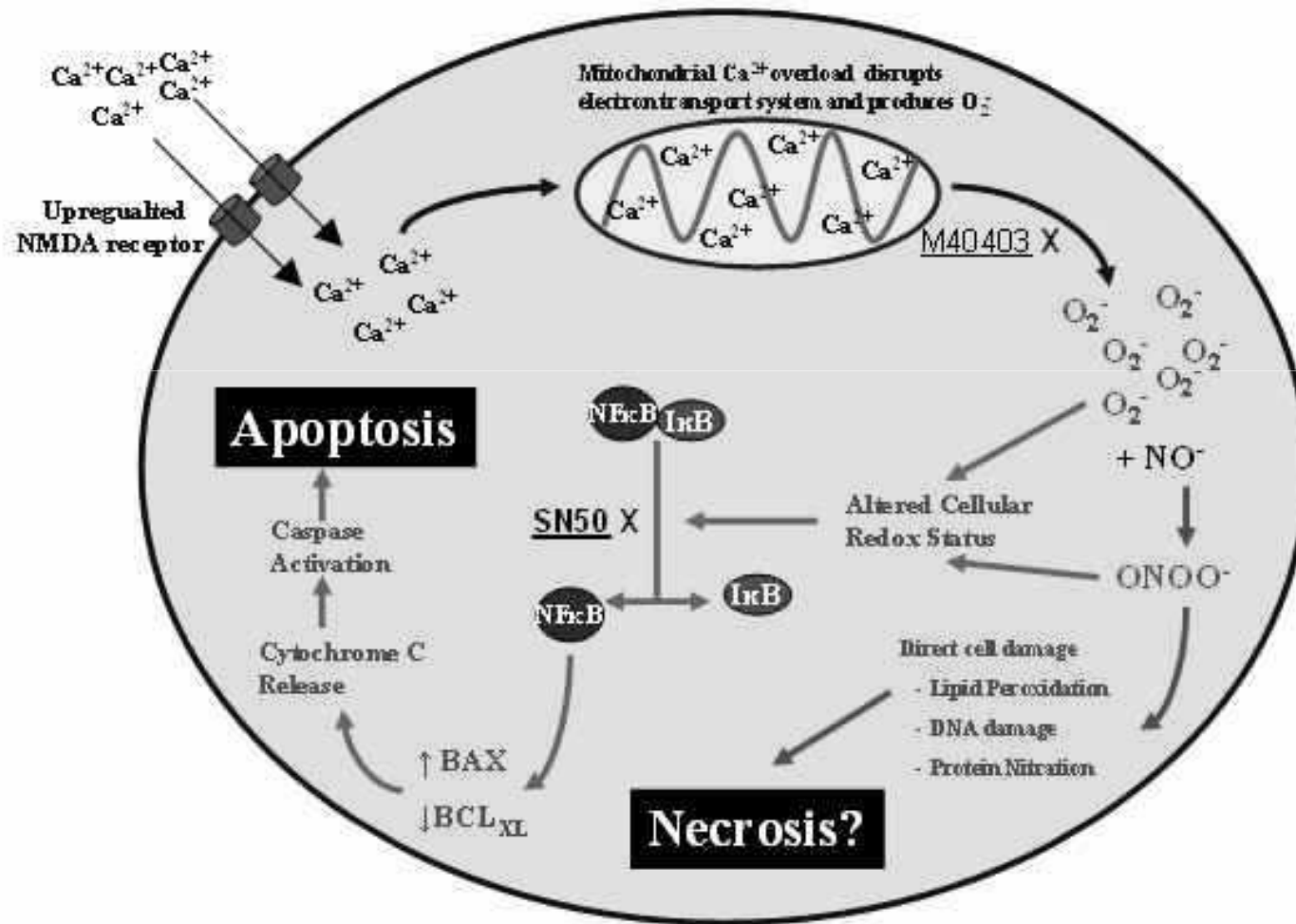
November 30, 2006



Ubiq=ubiquitous; Mat = maternal; activ = activator; rep = repressor;
unkn = unknown; Nucl. = nuclearization; X = β-catenin source;
nβ-TCF = nuclearized β-catenin-Tcf1; ES = early signal;
ECNS = early cytoplasmic nuclearization system; Zyg. N. = zygotic Notch

Copyright © 2001-2006 Hamid Bolouri and Eric Davidson

A system response



Apoptosis: programmed cell death

Necrosis: accidental cell death

Neuroinformatics

- Understanding the human nervous system is one of the greatest challenges of 21st century science.
- Its abilities dwarf any man-made system - perception, decision-making, cognition and reasoning.
- Neuroinformatics spans many scientific disciplines - from molecular biology to anthropology.

Neuroinformatics

- *Main research question:* How does the brain and nervous system work?
- *Main research activity:* gathering neuroscience data, knowledge and developing computational models and analytical tools for the integration and analysis of experimental data, leading to improvements in existing theories about the nervous system and brain.
- *Results for the clinic:* Neuroinformatics provides tools, databases, models, networks technologies and models for clinical and research purposes in the neuroscience community and related fields.

