Project no. 043338

Project acronym: EMERGENCE

Project title: A foundation for Synthetic Biology in Europe

Instrument: NEST Pathfinder

Thematic Priority: Synthetic Biology

Deliverable 2.2: Report on the possibilities and feasibility of implementing a European Master in Synthetic Biology

Due date of deliverable: August 2007
Actual submission date: October 2007

Start date of project: 1.12.2006 Duration: 36 months

Alfonso Jaramillo
Laboratoire de BIOCHIMIE. CNRS UMR7654 Ecole Polytechnique Route de Saclay91128 PALAISEAU Cedex Tel: +33-1-69334861 - FAX: +33-1-69334909

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	x
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

PROPOSAL FOR A EUROPEAN MASTER IN SYSTEMS AND SYNTHETIC BIOLOGY (mSSB)

Alfonso Jaramillo and François Képès

October 13th, 2007

GLOBAL STRUCTURE

The European Master 2 mSSB is composed of three optional Upgrade and ten compulsory block-courses (see Table). All these courses are planned during the first quarter of the Master 2 year. The two remaining quarters are devoted to rotations in laboratories.

The three optional block-courses are grouped into an "Optional Upgrade" Module. They provide introductions to Biology (one week), Computer science (half week) and Mathematics (half week). They are designed to enable acceptance of excellent students who would be slightly below threshold with respect to a small number of prerequisites, given the strong multidisciplinarity of the Master.

This is followed by a day of general introduction (not visible on the Table). Two Senior Full Professors will give the students a general scientific view of the Master and link its components in both bottom-up and top-down fashions. One MIT Professor will provide his vision of Synthetic Biology. The Director of iGEM (MIT) will introduce this international competition. Finally, the coordinator of the Master will provide informations on the organization of courses and exams.

The ten compulsory block-courses then start. They are generally distributed over several weeks, except when regrouping is necessary for organizational reasons (practical sessions) or pending the availability of non-local Professors.

The ten compulsory block-courses are distributed into four Modules. The first one, "Bio-inspiration" contains a single block-course. The three other Modules each comprise three block-courses: "Systems Biology", "Systems & Synthetic Biology", and "Synthetic Biology".

master of Systems and Synthetic Biology 2008	*CHAIR*	*CO-CHAIR*	*OTHER PROFESSORS*
Introduction to genomical Biology	F. Képès	F. Quétier	F. Toma, E. Westhof, M. D
Fundamental concepts of Computer Science	H. Klaudel		G. Hützler, E. Angel
Introduction to Mathematics for Biology	B. Prum	C. Ambroise	
An integrated and spatial view of the cellular machinery: from biology to modelling	H. Hirt	A. Paldi	J. Hérisson, J.-P. Renou, S
Integrated modelling for Physiology	S.R. Thomas		A. Hernandez, G. Hützler
Statistical analysis of biological sequences and gene expression	B. Prum	C. Ambroise	
Introduction to machine learning for network inference	F. d'Alché-Buc		F. Zehraoui, N. Brunel
Molecular modelling: protein interactions and protein design	D. Borgis \ A. Jaramillo	N. Basdevant	
Symbolic approaches to genetic regulatory networks	M. Aiguier [ECP]	P. Le Gall	
Language and modelling for design in systems and synthetic biology	F. Delaplace	H. Klaudel	O. Michel
Design, construction and characterization of biological parts and devices	A. Jaramillo	F. Molina	
Practice of genetic engineering	P. Dupuis-Williams	F. Képès	
Modelling and engineering networks of molecular interactions	P. Marlière	A. Jaramillo	S. Panke, V. Schächter

MODULES
Optional upgrade
Bio-inspiration
Systems Biology
Systems & Synthetic Biology
Synthetic Biology

[ECP] : École Centrale de Paris

Global structure of the master 2 mSSB. *One line corresponds to one block-course. The structure of the Table does not necessarily reflects the order of the courses, but provides a global view of the project main lines.*

PEDAGOGICAL CONTENT

"Upgrade" Module

Title
INTRODUCTION TO GENOMICAL BIOLOGY (optional)

Chair
François Képès

Co-chair
Francis Quétier

Other Professors
Flavio Toma, Éric Westhof, Frédéric Dardel, Marie Dutreix, Ivan Matic, Nadine Peyrieras, Jean-Marc Verbavatz, Hervé Delacroix, Alfonso Jaramillo

Prerequisite
None.

Synopsis
To facilitate research at the common borders, an Advanced Introduction aims at fostering a better understanding of the expectations, constraints, approaches and mode of thinking of a scientific partner across disciplines. Within a week, an AI brings non-biologists from a null/medium level to some understanding of the research frontiers in the biological sciences. To reach this goal, seasoned lecturers present the key objects and concepts of the target domain, explain the current research questions and methods, and give a feel for what would be considered a (good) result. All the important and recent subdisciplines of biology will be covered. The AIs have been successful since 2003, and at a European level since 2006.

Content
Molecular Genetics
Structural Biology
Cell Biology
Evolution
Developmental Biology
Genomics
Biological Networks
Synthetic Biology

Exam
Oral presentation with slides and critical discussion of a scientific paper chosen by the jury.

References
Dardel, F. and Képès, F. (2006). "Bioinformatics: Genomics and post-genomics". Wiley, UK (241 pages). ISBN 0-470-02001-6.

Title
FUNDAMENTAL CONCEPTS OF COMPUTER SCIENCE (optional)

Chair
Hanna Klaudel

Other Professors
Guillaume Hützler, Eric Angel

Prerequisite
None.

Synopsis
The aim of this lecture is to introduce fundamental notions of computer science for modellers. It covers three complementary fields including essential theoretical tools necessary to approach computer science aspects treated in this Master programme. The pedagogic volume is one half-week.

Content
1. Logic
    - propositional logic
    - predicate logic
2. Formal languages
    - regular expressions
    - finite automata
    - grammars
3. Algorithmics, calculability and complexity
    - standard data structures and algorithms (walks): lists, trees, graphs
    - decision problems, calculability
    - algorithmic complexity, complexity classes

Exam
None.

References
Hopcroft John E., Jeffrey D. Ullmann (1979) Introduction to Automata Theory, Language and Computation, (Addison-Wesley).
Michael Huth, Mark Ryan (2004) Logic in Computer Science: Modelling and Reasoning about Systems (Cambridge University Press), 2nd edition. ISBN 0-521-54310-X.
Alfred V. Aho, John E. Hopcroft, Jeffrey D. Ullman (1983) Data Structures and Algorithms (Addison-Wesley), ISBN 0201000237.

Title
INTRODUCTION TO MATHEMATICS FOR BIOLOGY (optional)

Chair
Bernard Prum

Co-chair
Christophe Ambroise

Prerequisite
Basic level in mathematics : elementary calculus, bases of linear algebra (resolution of linear equations, matrices), bases of probability theory (manipulation of probabilities of events, usual distributions : binomial, Poisson, gaussian, ...)

Synopsis
Today, Biology, and in particular Systems and Synthetic Biology, deals with a huge amount of data (in most cases very heterogeneous data) and with more and more complex modelizations. To analyze the experimental results, to choose optimal models for these results and estimate the parameters, as well as to describe and predict the behaviour of biological systems, mathematical tools are essential.
Such mathematical approaches will be used – and will therefore be introduced – in each following Block-Courses. This introductory course will present the necessary basic mathematical knowledge. The pedagogic volume is one half-week.

Content
Differential Equation (ODE) : problematics, theoretical resolution in simple cases, including the linear equations (essentially of order 1, but eventually dealing with several variables). Numerical resolution of ODE (Runge-Kutta);

Statistics : how the statistical reasoning differs from the general mathematical one?
Estimation;
Classical Hypothesis Testing;
Estimation by confidence intervals.

Exam
None.

References
Nuel, G. and Prum, B. (2007) Analyse Statistique des Séquences Biologiques (Hermes Sciences).

"**Synthetic Biology" Module**

Title

DESIGN, CONSTRUCTION AND CHARACTERIZATION OF BIOLOGICAL PARTS AND DEVICES

Chair

Alfonso Jaramillo

Co-chair

Franck Molina

Prerequisite

Some experience with bioinformatics, molecular and structural biology would be advantageous but not necessary.

Synopsis

The objective is to introduce the engineering principles of synthetic biology and to provide the methodology to engineer a cell with a targeted function using off-the-shelf biological parts. We will use as test case studies applications in medicine, biofuel, cellular biosensors and bioremediation.

Content

1. Introduction to Synthetic Biology
   - Design principles (decoupling, abstraction, standardization)
   - Comparison of the classical and abstracted vision of biological components and functions
   - Concept of biological parts, devices and chassis. Registry of Parts.
   - Current main and alternative strategies to build synthetic biological parts or systems
2. Biological parts
   - Catalogue of parts
   - Methods for designing parts
   - Construction of parts
   - Characterisation of parts
3. Biological devices
   - Catalogue of devices
   - Methods for designing devices
   - Construction of devices
   - Characterisation of devices
   - Debugging of devices
4. Biological systems
   - Design and construction of a chassis
   - Use of –omics measurements for systems characterisation
   - Applications
5. New directions and expected progresses

Exam

Write up and oral defense of a project to design a system from parts.

References

D Ferber (2004). Microbes Made to Order. Science 303, 158 - 161
J Hasty, et al. (2002). Engineered Gene Circuits. Nature 420, 224 - 230.
MB Elowitz and S Leibler (2000). A Synthetic Oscillatory Network of Transcriptional Regulators. Nature 403, 335-338.

S Basu et al. (2005). A synthetic multicellular system for programmed pattern formation. Nature 434, 1130-1134.

Title

LANGUAGE AND MODELLING FOR DESIGN IN SYSTEMS AND SYNTHETIC BIOLOGY

Chair

Franck Delaplace

Co-chair

Hanna Klaudel

Other Professors

Olivier Michel

Prerequisite

Basics knowledge of programming languages.

Synopsis

The design of synthetic biological functions relies on general principles originating from engineering like decomposition, abstraction and standardization. The current realizations tend towards the application of these principles, some parts of which being performed by software. They conceptually follow a scheme close to the compiling chain (i.e., data-processing tools carrying out the translation from a specification defined by designers towards an analysable representation that can be finally processed by a computer). According to this leading framework, the goal of this lecture is to present a panorama of the fundamental computer science methods with a strong emphasis to systemic and synthetic biology topics. It starts from a description in a specification language/formalism, translates it in an analysable model in order to validate bio-safety and bio-security aspects; and finally ends by the code generation stage producing the final assembly carrying out a biological function. The integration of the methods will pragmatically address case studies inspired from some iGEM realizations.

Content

1. Languages & Formalisms
   - Modelling and Simulation (rewriting based languages, amorphous computing)
   - Analysis and verification (process algebra: links to Pi-calculus and model-checking techniques)
2. Models & Analysis
   - Game theory
   - Petri Nets
3. Application

Exam

Written exam.

References

C. Chettaoui, F. Delaplace, P. Lescanne, M. Vestergaard & R. Vestergaard (2006) Rewriting Game Theory as a Foundation for State-Based Models of Gene Regulation. *In* International conference on Computational Methods In Systems Biology (CMSB).

A. Yartseva, H. Klaudel, R. Devillers, F. Képès (2007) Incremental and unifying modelling formalism for biological interaction networks. BMC-Bioinformatics.

J.-L. Giavitto, G. Malcolm, O. Michel (2004) Rewriting systems and the modelling of biological systems. Comparative and Functional Genomics 5, 95-99.

Djebali, S., Delaplace, F., Roest Crollius, H. (2006) Exogean : a framework for annotating protein-coding genes in eukaryotic genomic DNA. Genome Biology 7 (Suppl 1):S7.

"Systems and Synthetic Biology" Module

Title

MOLECULAR MODELLING: PROTEIN INTERACTIONS AND PROTEIN DESIGN

Chairs

Daniel Borgis, Alfonso Jaramillo

Co-chair

Nathalie Basdevant

Prerequisite

Basic notions of: molecular and structural biology, bioinformatics, point mechanics, thermodynamics, differential equations.

Synopsis

This teaching unit presents a self-contained introduction to biomolecular modelling, either at a traditional (atomic-scale) level of description, or at a coarse-grained level. The basic unifying principle is the definition of a force field between elementary constituents that obeys the fundamental laws of physico-chemistry. From a systems biology point of view, beyond the traditional problem of the prediction of biomolecules structure, dynamics, and reactivity, the main goal is the in-silico description of biomolecular self-assemblies of increasing complexity, starting with the problem of protein-protein and protein-DNA docking. From a synthetic point of view, the emphasis will be put on computational methods for protein design and the computer-aided design of new biological parts such as biosensors, novel enzymes and thermostable proteins, specific molecule, DNA, or protein binding sites.

Content

1. Structural biology: structure of proteins, nucleic acids, and their complexes
2. Molecular Modelling principles: atomic models and coarse-grained models, potential energy function, and solvent representation
3. Energy minimisation methods
4. Molecular Dynamics Simulation Methods: principles and algorithmic tools, analysis tools
5. Monte-Carlo Simulation Methods
6. Brownian Dynamics: brownian movement, stochastic dynamics
7. Biological applications: interaction free energy, protein-protein docking
8. Computational protein design using atomic models: folding and inverse folding, rotamer library, protein folding energy, combinatorial optimization methods
9. Applications: redesign and de novo design of small molecules to enzymes.

Exam

Article study and writing of a report with a brief research project. For example, design of a protein with a novel function.

References

A. R. Leach (2001) Molecular Modelling. Principles and Applications. (Pearson Education, England; 2nd edition).

V. Tozzini. Coarse-grained models for proteins (2005) Curr. Opin. Struc. Biol. 15: 144 -150.

N. Basdevant, D. Borgis & T. Ha-Duong. (2007) A Coarse-Grained Protein-Protein Potential Derived from an All-Atom Force Field. J. Phys. Chem. B. 111: 9390-9399.

B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard & D. Baker. (2004) Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. Science 302: 5649.

A. Jaramillo, L. Wernisch, S. Henry, and S. Wodak. (2002) Folding free energy function selects native-like protein sequences in the core but not on the surface. PNAS 99:13554.

"Bio-inspiration" Module

Title
AN INTEGRATED AND SPATIAL VIEW OF THE CELLULAR MACHINERY: FROM BIOLOGY TO MODELLING

Chair
Heribert Hirt

Co-chair
Andras Paldi

Other Professors
Joan Hérisson, Jean-Pierre Renou, Sébastien Aubourg, Delphine Pflieger

Prerequisite
Notions of genomics, transcriptomics, proteomics, metabolomics.

Content
Vision and challenges of the future
Functional organization and structure of the nucleus
Chromatin and chromosomes, chromatin dynamics and Epigenetics
Biophysical models of chromosomes
Biological model systems: From bacterial to eukaryotic signal transduction
An integrated view of regulatory processes:
- Unraveling the genome information
- Transcriptional control networks
- Protein Interacting networks

Synopsis
The course should provide an insight into our current genomic and epigenomic concepts: Moreover, it should provide the student with an understanding how information is sensed and processed by biological systems and how genomic and epigenomic responses are linked to these processes. An outlook into future challenges and applications will be given, demonstrating the usefulness of these studies. Finally, the course is intended to link genomics to the other modules in the masters programme with the aim to stimulate interdisciplinary approaches and thinking.

Exam
Students will be evaluated on the basis of their ability to give a seminar on one of the topics.

References
Nielsen, HB, Mundy, J, Willenbrock, H (2007) FARO: Functional Associations by Response Overlap, a functional genomics approach matching gene expression phenotypes.PLoS ONE 2: 676.
Plavec I, Sirenko O, Privat S, Wang Y, Dajee M, Melrose J, Nakao B, Hytopoulos E, Berg EL, Butcher EC. (2004) Method for analyzing signaling networks in complex cellular systems. Proc Natl Acad Sci U S A. 101:1223-8
Mathe C, Sagot MF, Schiex T, Rouze P. (2002) Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res. 30:4103-17.

Title
SYMBOLIC APPROACHES TO GENETIC REGULATORY NETWORKS
École Centrale de Paris

Chair
Marc Aiguier

Co-chair
Pascale Le Gall

Prerequisite
Notions in computer sciences: algorithms, graphs, logic.

Synopsis
The functioning of a cell is controlled by large interaction networks between genes, proteins and other molecules, called Regulatory Networks. To describe the dynamics of Genetic Regulatory Networks (GRN), several differential equation systems have been proposed, but they depend on parameters which are often unknown due to lack of observations. Understanding the functioning of GRN supposes a modelling of biological processes representing the set of all possible behaviours or models on which reasoning becomes possible. Boolean or multi-valued models have been introduced to describe the qualitative features of the dynamics. Formal methods issued from Software Engineering (SE) have been successfully applied in this context: model-checking allows the verification of temporal properties of the models of GRN. The following questions can then be addressed: is there a model coherent with a certain observed temporal property? Are there some temporal properties compatible with all possible models? Symbolic SE techniques allow one to directly handle sets of models and thus, increase model-checking efficiency. Moreover if the structure of GRN can be seen as interconnecting sub GRN, each one modelling a biological function, then a GRN can be compared to a complex system made of pieces of software.

Content
1. Introduction to Genetic Regulatory Networks (GRN)
   - From differential equation systems to discrete ones.
   - Reasoning on dynamics of GRN: validating and predicting biological knowledge about GRN by analysing models
2. Automation of the analysis of discrete models with temporal logics
   - Model-checking and bisimulation
   - Symbolic Execution and constraint programming
   - Illustration example: mucoidy regulation in *Pseudomonas aeruginosa*
3. Structure of GRN as several interconnecting sub GRN
   - The question of inherited properties from sub GRN
   - Open research questions: Folding of stationary circuits, new GRN combination connectors.

Exam
A final 3-hour exam at the end of the course.

References
Clarke E.M., Grumberg P. and Peled D.A. (1999) Model Checking (MIT Press).
Thomas R., Thieffry D. and Kaufman M. (1995) Dynamical behaviour of biological regulatory networks-I. Biological role of feedback loops and practical use of the concept of the loop-characteristic state. Bull Math Biol. 57, 247-276.

"Systems Biology" Module

Title
INTEGRATED MODELLING FOR PHYSIOLOGY

Chair
S. Randall Thomas

Other Professors
Alfredo Hernandez, Guillaume Hützler

Prerequisites
Basic Ordinary Differential Equations (ODE); Introductory Biology.

Synopsis
This course will give an integrated view of multi-organ physiological regulation, with blood pressure regulation taken as a detailed example. We will present the multi-resolution modeling environment called SAPHIR, a prototype "core-model" of the International Physiome project, which is based on classic Guyton models, and extended to treat organism-level implications of single-gene polymorphisms. As a second example, using the Multi-Agent modelling technique, the Epitheliome project will be presented. As a practical complement to the course, the students will use a user-friendly ODE-solver (Berkeley Madonna) to work through sample problems in multi-compartmental analysis and nonlinear pharmacokinetics.

Content
Vision/Introduction
Notions of Irreversible Thermodynamics, coupled fluxes and membrane-bound compartments
Physiology of hypertension, systems engineering approach; modelling - the SAPHIR physiome model
Epitheliome: an example of Multi-Agent modeling in epithelial physiology
Computational aspects (including one-week supervised personal work with "Berkeley Madonna")
Clinical aspects: Parameter identification of cardiac models for individual patients

Exam
Evaluation of hands-on solutions to a set of problems with Berkeley Madonna.

References
Guyton, A. C., T. G. Coleman and H. J. Granger (1972). Circulation: overall regulation. Annu Rev Physiol 34: 13-46.

Title
PRACTICE OF GENETIC ENGINEERING

Chair
Pascale Dupuis-Williams

Co-chair
François Képès

Prerequisite
None.

Synopsis
One of the virtues of the concept of Synthetic Biology is to lower for non-biologists the ticket price for getting their hands "wet", ie to practice genetic engineering themselves. This is why practical works in this technology are an integral and central part of any Master degree that covers Synthetic Biology. This block-course has been designed with this mindset, in a strong interaction with the previous one, "Design, construction and characterization of biological parts and devices".

Objective

The practical works are aimed at providing the future engineers and scientists with knowledge of the fundamental principles and the classical methodologies of genetic engineering. The purpose of this practical work is to provide a theoretical and experimental basis for the synthetic biology programs involving cloning, both in fundamental research (conception of prokaryotic systems for the study or control of gene expression) or in engineering (production or modification of chemical and biological substances for the health sector, the food processing industry, energy production or the treatment of pollutants).

The organisation of the practical work allows 42 hours for a sequence of experiments which will introduce students to the notions of biological experimental strategies and processes and encourage them to analyse the potentials and limits of biological systems *in vitro* and *in vivo* by appropriately assessing the variables and productivity of the experiments.

Content
Theoretical contents

At the theoretical level, the practical work illustrates the fundamental principals of molecular biology: DNA structure, bacterial operons, strategies for regulated gene expression, protein production using genetic engineering, analytical methods of biomolecules (nucleic acids and proteins). It also illustrates the complexity of living organisms and analyses the levels of control over processes in integrated *in vitro* systems (enzymatic systems) and *in vivo* systems (micro organisms).

Methodological contents

The classical methodologies of DNA cloning in bacteria will be used, in particular: physico-chemistry of DNA, PCR amplification, use of expression plasmids, bacterial transformation and clonal selection via antibiotics, as well as induced expression, analysis and purification of bacterial proteins.

Pedagogic contents

The practical work sessions are designed to progressively introduce methodological and strategic « alternatives », allowing students to develop their autonomy and their innovative spirit.

Detailed contents

The first part (3 days) is spent on the construction of the gene of interest and its insertion into a bacterial vector.

The 2nd part (2 days) is used for the bacterial transformation and the phenotypic and genetic selection of the clones.

The 3rd part (2 days) is dedicated to the induced expression, purification and analysis of bacterial proteins.

Exam
The quality of the manipulations will be evaluated during the practical work.

Title
MODELLING AND ENGINEERING NETWORKS OF MOLECULAR INTERACTIONS

Chair
Philippe Marlière

Co-chair
Alfonso Jaramillo

Other Professors
Vincent Schachter, Sven Panke

Prerequisite
Bases of molecular and cell biology. Some experience with bioinformatics and kinetic models for chemical reactions would be advantageous but not necessary.

Synopsis
The objective of this course is to introduce modelling frameworks for metabolic and transcriptional networks, their use for metabolic engineering and the corresponding industrial applications.
Modelling of metabolic and regulatory networks is one of the pillars of both systems and synthetic biology. The first part of this course will review several modelling frameworks, together with applications typical of each, including phenotype prediction, experimental data integration, reverse engineering of regulatory influences, or analysis of simple dynamical properties.
The second part of the course will focus more specifically on metabolic modelling and one of its core application, metabolic engineering. Metabolic engineering has emerged in the past 15 years as an interdisciplinary field aiming at the improvement of cellular properties by using modern genetic tools to modify pathways. Recent advances in DNA synthesis, genome engineering, high-throughput analytics, adaptive evolution, and model-based analysis of biochemical systems and protein engineering have expanded the field towards the redesign of reaction systems of significant complexity within cells. We will review these advances and their impact, using examples of recent successful metabolic engineering strategies.

Content
1. Introduction to metabolic modelling
2. Genome wide reconstruction of metabolic and genetic regulation networks
    - Integrating high-throughput experimental data within models
3. Metabolic engineering of microbial strains
    - Engineering of metabolic pathways for optimized production of chemicals
4. Computational design of networks of molecular interactions
5. Industrial applications of metabolic engineering.

Exam
Write up and oral defence of a project to study or design a biological network using available experimental data.

References
Barrett, C.L. et al. (2006) Systems biology as a foundation for genome-scale synthetic biology. Curr Opin Biotechnol. 17, 488-92.
Di Ventura B. et al. (2006) From in vivo to in silico biology and back. Nature 443, 527-33.

Endy D. (2005) Foundations for engineering biology. Nature 438, 449-53.
Pfleger B.F. et al. (2006) Combinatorial engineering of intergenic regions in operons tunes expression of multiple genes. Nat Biotechnol. 24, 1027-32.
Sprinzak D. & Elowitz M.B. (2005) Reconstruction of genetic circuits. Nature 438, 443-8.

Title
INTRODUCTION TO MACHINE LEARNING FOR NETWORK INFERENCE

Chair
Florence d'Alché-Buc

Other Professors
Farida Zehraoui, Nicolas Brunel

Prerequisite
A background in statistics, probability theory, optimization and algorithmic theory (graph theory).

Synopsis
Recent developments on statistical machine learning applied to structured data have led to important improvements in the field of biological network inference. The purpose of this introductory course is to familiarize master students with the concepts and algorithms of network inference from data in the context of concrete applications in Systems Biology and Synthetic Biology: network reverse-engineering, model parameter estimation, network completion, protein function prediction. The course will provide an overview of unsupervised and supervised approaches devoted to network inference (resp. Part 2 and Part 3). Part 2 is strongly related to the course "Analysis of gene expression" where other unsupervised methods will be described, some of which being devoted to graph mixtures and thus very relevant for network inference. This part which is also concerned with Continuous Markov Models such as state-space models is also related to the course about Markov models for sequence analysis. Part 3 is related to novel tools developed recently around prediction of structured outputs in machine learning and information retrieval.

Content
1. Motivation for biological network inference: a short overview
    - Brief introduction to statistical theory of machine learning
2. Unsupervised approaches to network inference
    - Models for reverse-engineering (regulatory networks, signalling pathways, metabolic networks)
    - Parameters and structure estimation
    - Examples: signalling pathways, regulatory, metabolic networks (distributed over the course)
3. Supervised approaches to network inference (protein-protein networks, metabolic pathways)
    - Supervised approaches
    - Applications to protein-protein networks, enzyme networks and regulatory networks (distributed over the course)

Exam
The students will carry out a project (litterature analysis or test of some extraction algorithm on a dataset) and will have a written exam.

References
Gardner T.-S., di Bernardo D., Lorenz D., Collins J.J. (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. Science:301:102–105.

Perrin B.-E., Ralaivola L., Mazurie A., Bottani S., Mallet J., d'Alché-Buc F. (2003) Gene networks inference using dynamic Bayesian networks. Bioinformatics,19:II138–II148.
Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.Nat. Genet., 34(2):166-76.
Yamanishi, Y., Vert, J.-P. and Kanehisa, M. (2005) Supervised Enzyme Network Inference from the Integration of Genomic Data and Chemical Information. Bioinformatics, Vol.21:i468-i477.

Title
STATISTICAL ANALYSIS OF BIOLOGICAL SEQUENCES AND GENE EXPRESSION

Chair
Bernard Prum

Co-chair
Christophe Ambroise

Prerequisite
Elementary statistics: Pointwise estimation, classical hypothesis testing and confidence intervals. Basics of molecular biology: protein, DNA, RNA, transcription, translation.

Synopsis
The large amount of available biological sequences and gene expression data have changed the way biologists experiment to explore the genome. The analysis of such data requires knowledge in exploratory data analysis (visualization, statistic summary), in modeling dependence and in inferential statistic.
This course will present the necessary statistical knowledge to analyze both types of data. It is divided in two parts. The first part is dedicated to biological sequence analysis using Markov models. The second part presents the main methods in exploratory data analysis (unsupervised learning) and inferential statistics required for exploiting microarrays. This last family of methods will also be of use for the analysis of biological networks.

Content
Part 1: Biological sequence analysis

- Markov models for biological sequence analysis (DNA and proteins) : One of the main goal of this statistical modelling is the search for exceptional motif. Exceptionality may be related to a specific biological function.
- Hidden Markov Model : Modelling the sequence using different regimes which corresponds to homogeneous parts of the sequence allows to better adjust to the underlying biological reality (i.e. introns / exons / coding/non-coding). Moreover identifying a hidden Markov model allows to segment the sequence and is thus helpful for annotation.
- Profiles-HMM : Searching for profiles along the sequence (statistical distribution of nucleotides or amino acids) allows to localize biological signals (i.e. ``binding sites", etc.). Classical approach and automate based approach (highly efficient) will both be presented.

Part 2: Statistical Analysis of Gene Expression

- Visualization: Descriptive Statistics, Principal Component Analysis, Multidimensional Scaling.
- Multiple Hypothesis Testing : Type I error, Strong and weak control of the error. Controlling FWER, Bonferroni and Sidak, descending methods. Controlling FDR, ascending method of Benjamini and Hochberg. Non parametric testing, permutation tests (illustrating with SAM).
- Clustering: Structures, goal and methodology of clustering. Statistical approaches for clustering. Parametric models, non parametric models, mixture models.

Exam

References
McLachlan, G., Do, K. and Ambroise, C. Analyzing microarray gene expression data (Wiley).
Nuel, G. and Prum, B. (2007) Analyse Statistique des Séquences Biologiques (Hermes Sciences).