

# Tutorial Proposal for ISMB05

## Title

Weighted Finite-State Transducers in Computational Biology

## Instructors

Corinna Cortes  
Google Research

Mehryar Mohri  
Courant Institute, NYU

Corinna Cortes is a Research Scientist at Google, Inc. where she is working on a broad range of theoretical and applied large-scale machine learning problems. Prior to Google, Dr. Cortes spent more than ten years at AT&T Labs - Research, formerly AT&T Bell Labs, where she held a distinguished research position. Dr. Cortes' research work is well-known in particular for her contributions to data-mining in very large data sets for which she was awarded the AT&T Science and Technology Medal in the year 2000, and her work on the theoretical foundations of support vector machines (SVMs) and kernel techniques for the analysis of variable-length sequences and weighted automata. She has been giving numerous talks and presentations in machine learning.

Mehryar Mohri is a Professor of Computer Science at the Courant Institute of Mathematical Sciences. Before joining NYU, Dr. Mohri worked for ten years at AT&T Bell Labs and AT&T Labs - Research where he served as the Head of a Research Department, leading and directly contributing to a broad range of work in machine learning, automata theory, text and speech processing, and the design of general algorithms. He is co-author of the software libraries FSM Library and GRM Library used by thousands in research centers and universities across the world. He has taught dozens of tutorials on topics related to weighted finite-state transducers, their theory and algorithms, and their applications to sequence modeling.

## Motivation and Goals

Projects such as genome sequencing and DNA microarray studies produce an ever-increasing amount of data and the area of computational biology now poses some of the biggest challenges in computer science and data mining such as data storage, visualization, and modeling. Statistical learning techniques are increasingly successfully applied for modeling, but they often require substantial algorithmic expertise: the conventional software packages do not naturally encompass variable-length sequences and general structures, thus, special-purpose algorithms must be implemented to solve the associated optimization problems. In this tutorial, we introduce a general framework, weighted finite-state transducers (WFSTs), that naturally matches a wealth of data in computational biology, together with a set of algorithms that efficiently allow for a series of operations on WFSTs, enabling biologists even with modest computer skills to successfully apply sophisticated alignment techniques and powerful kernel methods.

There has been a substantial amount of progress in the development of the theory and the design of algorithms based on WFSTs over the last two decades. General and efficient algorithms have been devised for combining and optimizing

WFSTs. These algorithms have been used in a variety of text, speech, and sequence processing applications where automata and transducers of several hundred million states and transitions are used.

The main objective of this tutorial is to familiarize the audience with WFSTs algorithms and techniques, their use and application to computational biology problems, and a software library (FSM Library) incorporating these algorithms and representations that is freely available for research and academic use.

The sequence modeling and algorithmic problems arising in computational biology can greatly benefit from the use of WFSTs and these algorithms. For example, rather than devising a new and distinct dynamic programming algorithm for each new type of pairwise or multiple sequence alignment problem, a *single* general composition algorithm can be used to efficiently compute all such alignments. The graphical representation of WFSTs can help one design new and more complex alignment models conveniently. General algorithms such as rational operations can be used to combine simpler models to create more elaborate ones. Many of the modern machine learning problems related to biological sequences such as classification, regression, clustering, can also be elegantly and effectively solved using kernel methods based on WFSTs.

This tutorial will survey their use for modeling and algorithmic problems in computational biology and point out how they can provide a common and natural representation for pairwise sequence alignment models, inexact matching, and the design and computation of general sequence kernels crucial for a variety of machine learning problems.

### **Intended Audience**

This tutorial is meant for a broad audience. Students, researchers, biologists and computer scientists, interested in an overview of general and efficient algorithms for statistical learning techniques used in computational biology, in the design and computation of alignment models, sequence kernels for the analysis of problems such as the remote homology problem, and many other sequence learning and modeling problems such as clustering, regression, ranking problems. No specific knowledge will be required since the tutorial is self-contained and most fundamental concepts are introduced during the course. The tutorial will also familiarize the audience with a software library implementing the above mentioned algorithms but no knowledge of programming languages is required. Some basic knowledge of algorithms and bioinformatics will be helpful.

### **Tutorial level**

Introductory level.

### **Detailed Outline**

1. Weighted-finite state transducer algorithms and software library (FSM Library)
  - Introduction
  - Acceptors/Transducers
  - Weights/Semirings

- Rational Operations - Union, Concatenation, Kleene Closure
  - Product Operations - Intersection, Composition, Difference
  - Relabeling Operations - Projection, Inversion, Reversal
  - Search Operations - Best Path, N-Best Path, Pruning
  - Equivalence Transformations - Connectivity, Epsilon-Removal, Determinization, Weight and Label Pushing, Minimization
2. Pairwise alignments for bioinformatics: representation and computation with weighted transducers
- Definitions and biological motivation
  - Unified framework
  - Single general computation algorithm
  - Alignment transducers
  - Representation and computation with the FSM library
  - Global alignment transducers
  - Local alignment transducers
  - Gappy alignment transducers
  - Complex alignment transducers
  - Learning alignment transducers
  - Alignment of sets of sequences
  - Alignment of automata
  - Longest common subsequences of automata
3. Kernels for computational biology
- Motivation
  - Learning from examples
  - Overfitting
  - Optimal hyperplane
  - Support vector machines (SVMs)
  - Classification
  - Soft-margin classifier
  - Kernels
  - Mercer's condition
  - Unified transducer framework for string kernels
  - Sequence kernels for bioinformatics - spectrum/mismatch kernels, convolutional kernels, gappy kernels, edit-distance kernels
  - Representation with weighted transducers
  - Computation algorithm
  - Applications in computational biology - the recognition of translation initiation sites, remote protein homology problem, analysis of phylogenetic profiles, prediction of signal peptide cleavage site, analysis of phylogenetic profiles
  - Other kernel techniques and applications - clustering, regression, ranking.