Ecole Doctorale STIC

Sciences & Technologies de l'Information et de la Communication

MINES ParisTech

Université Nice Sophia Antipolis

ED STIC - Proposition de Sujets de Thèse

pour la campagne d'Allocation de thèses 2011

Titre du sujet : Scoring and discriminating in high-dimensional spaces: a geometric based approach of statistical tests

Mention de thèse : Informatique

HDR Directeur de thèse inscrit à l'ED STIC : Cazals Frederic

Co-encadrant de thèse éventuel :

Nom :

Prénom :

Email :

Téléphone :

Email de contact pour ce sujet : frederic.cazals@inria.fr

Laboratoire d'accueil : INRIA

Description du sujet :

Voir description en Anglais.

English version:

CONTEXT

In science and engineering, scoring is the problem concerned with the identification of the best models amidst a large collection of putative models. One classical case is scoring for protein docking: given two isolated molecules, docking algorithms generate putative protein complexes, and scoring aims at identifying the complexes which best resemble to what Nature does. In its simplest form, scoring is based on the ranking of real values incorporating knowledge on the system. For complex situations, though, each individual is represented by a point in a high-dimensional space, and the ranking must take into account multiple criteria, which are possibly conflicting. Scoring can also be supervised, if some examples of positive and negative cases can be used to learn which features should be favored.

The scoring functions developed so far, in particular for docking, suffer from two drawbacks. First, the ranking provided is often not significant: in terms of statistical hypothesis testing (p-value analysis), the outcome is tantamount to that one would obtain by pure chance [Feliu2010]. Second, even when multiple criteria are used, the decision is often based on one-dimensional real values [Lon08].

GOALS

The goal of this PhD thesis will be to develop novel tools aiming at discriminating two populations in a high-dimensional space. The theoretical background will be that of so-called rank tests [Mann47] on the one hand, and of dimensionality reduction [Joh84] and geometric learning [Caz11] on the other hand. (Geometric learning is a body of work concerned with the inference of features of a set from a collection of sample points.)

On the theoretical side, the goal will be to develop novel statistical tests with improved discriminative power. These tests will be geared towards the analysis of high-dimensional point clouds but also high-dimensional dynamical systems. On the applied side, these tests will be used, in particular, in the context of the scoring experiment of CAPRI, the community wide contest for protein docking, see http://www.ebi.ac.uk/msd-srv/capri.

BACKGROUND

The PhD candidate should have a strong background in theoretical computer science or applied mathematics or biophysics or statistics / probability theory, and a genuine interest for (structural) biology.

BIBLIOGRAPHY

[Fel10] Feliu, E. and Oliva, B; How different from random are docking predictions when ranked by scoring functions?; Proteins: Structure, Function, and Bioinformatics, 2010

[Lond08] London, N. and Schueler-Furman, O.; Funnel Hunting in a Rough Terrain: Learning and Discriminating Native Energy Funnels; Structure, 16 (2), 2008

[Caz11] F. Cazals and D. Cohen-Steiner; Reconstructing 3D compact sets; Computational Geometry: Theory and Applications, 2011

[Joh84] W. Johnson and J. Lindenstrauss; Extensions of Lipschitz mappings into a Hilbert space; Contemporary Mathematics, 26, 1984.

[Mann47] Mann, H. B.and Whitney, D. R; On a Test of Whether one of Two Random Variables is

Stochastically Larger than the Other; Annals of Mathematical Statistics 18 (1), 1947.

MISC

-- the Algorithms-Biology-Structure group from INRIA Sophia-Antipolis-Méditerranée:
http://www-sop.inria.fr/abs

-- full description of the PhD thesis:
ftp://ftp-sop.inria.fr/abs/fcazals/positions/thesis11_scoring.pdf

URL : ftp://ftp-sop.inria.fr/abs/fcazals/positions/thesis11_scoring.pdf