# International Training Program (ITP)
# Participation Application Form

- **Applicant (Student) Name** : J.B. Brown
- **Home Institute**
  - Laboratory of Biological Network Information (Akutsu Laboratory)
    Kyoto University Institute for Chemical Research; Kyoto, Uji, Gokasho 611-0011
- **Partner Institute**
  - Macromolecular Modelling Group (Knapp Laboratory)
    Free University of Berlin; Fabeckstrassen 36a, D-14195, Germany, Berlin-Dahlem
- **Duration of program** : September, 2009 - November, 2009 (approx. 10 weeks)

## Research background and objectives of exchange program

### Background

In previous research, J.B. Brown has led a project for the development and testing of a computational technique for analyzing target properties in compounds. The approach was to design a measure of similarity between compounds, called a *kernel*, which is then utilized by a pattern analysis algorithm such as a Support Vector Machine. More specifically, the compound similarity technique is called a chiral graph kernel (labeled $K$), which measures the similarity of two compounds by transforming atoms and bonds into mathematical graphs $G_1, G_2$ and counting (via a function $\psi$) the number of pattern sub-structures (trees $t \in T$) existing in both compounds:

$$K(G_1, G_2) = \sum_{t \in T} w(t)\psi_t(G_1)\psi_t(G_2) \quad .$$

An example of this equation is shown below, where $w(t)$ is a weight given to each tree.

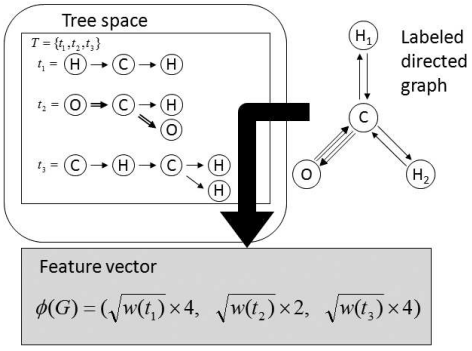


Figure 1: An example of transforming a chemical structure into a feature vector.

J.B. Brown, along with other collaborators, extended the tree patterns from previous research to include the chirality or handedness of molecules (e.g., *R*-ibuprofen and *S*-ibuprofen). As a result, it was shown that Support Vector Regression (SVR) with a chiral graph kernel produced a drastic improvement in target property prediction (Brown *et al.*, J. Chem. Inf. Modell., in re-review).

Despite the success of the kernel method proposed, it has yet to incorporate the physicochemical properties of atoms in a drug or other molecule. As a simple example, the compounds $CH_4$ and $SiH_4$ will have a similarity score of $K(CH_4, SiH_4) = 0$ because methane contains the pattern tree C→H, different from silane's pattern Si→H. However, it is known that similar valence group substitutions are important for industry and medicine.

## Research objectives

J.B. Brown has proposed two improvements to the existing chiral tree-pattern graph kernel to accomodate for compounds that interchange atoms of the same valence group. As both of the extensions are physicochemical, it is important that J.B. Brown collaborate with a research group with a focus on physical chemistry. As a result, J.B. Brown contacted professor Ernst Walter-Knapp of the partner institute, and J.B. Brown's visit to Knapp's group was agreed upon.

At the Free University of Berlin, J.B. Brown will collaborate with Knapp's group on the following projects:

1. Revision of the kernel equation given above to relax matching requirements to only valence group matchings. This is critical for compound toxicity analysis and virtual screening applications in pharmacology. In the previous method, trees of depth one (single atoms) were calculated for similarity as follows:

$$K(u, v) = \lambda * \mathbf{1}(label(u) = label(v)) \ ,$$

   where $\mathbf{1}()$ is a function returning 1 if a true argument is given, and 0 otherwise (lambda is a weighting of a match). The similarity of two trees is then the summation of these atom match weightings (convolution), and the similarity of two compounds is the similarity of the trees constituting the compounds (the equation on page 1). J.B. Brown's proposed extension, which will be argued and refined during the ITP program, is as follows:

$$K_e(u, v) = \mathbf{1}(|\delta^+(u)| = |\delta^+(v)|) * (\lambda * \hat{k}^{-|\mathrm{EN}(u) - \mathrm{EN}(v)|}) \ ,$$

   where we introduce the electronegativity $\mathrm{EN}(u)$ of an atom $u$, and an electronegativity match free parameter $\hat{k} > 1$. $\delta^+(u)$ returns the set of outgoing neighbors in compound $u$'s graph representation (see figure above).

2. Expansion of the kernel methods to three dimensional compounds with analysis via electrostatic potential (ESP). *ESP is the Knapp group's specialty, and therefore, the state of the art in kernel methods for biochemical QSAR analyses will be greatly enhanced as a result of this collaboration.* Knapp's group has the computing hardware, software, and field-specific knowledge to help J.B. Brown's QSAR research. J.B. Brown has proposed the following ESP kernel, though it has yet to be tested:

$$K_{\mathrm{ESP}}(C_1, C_2) = k_1 \exp^{-k_2(\Sigma_{p=(x,y,z)}|ESP(C_1^p) - ESP(C_2^p)|)} \ .$$

   Essentially, the equation above represents the exponential decrease in similarity by calculating the convoluted difference in electrostatic potential at each point in a virtual electrostatic mesh. *This means that analyses are no longer restricted to chemical compounds; biochemical complexes can now be evaluated and screened by kernel methods.* For systems and molecular bioinformatics, this will be a large advance. At Knapp's group, J.B. Brown will spend a good amount of time discussing how to set the constants $k_1$ and $k_2$ based on standard physicochemical properties.