Kernel methods in computational and systems biology

Jean-Philippe.Vert@mines.org

Ecole des Mines de Paris, France
Computational biology

The age of data in biology

And many more...

And many more...
Motivations

Develop a theoretical framework and algorithms in order to

• represent and integrate biological data

• model and conceptualize living systems

• infer properties of living systems
Biological data are often

- structured and heterogeneous: sequences, 3D structures, graphs, networks, expression profiles, phylogenetic trees, SNP, ...
Biological data are often

- **structured and heterogeneous**: sequences, 3D structures, graphs, networks, expression profiles, phylogenetic trees, SNP, ...

- in **large quantities** $(10^6$ gene sequences$)$
Biological data are often

- structured and heterogeneous: sequences, 3D structures, graphs, networks, expression profiles, phylogenetic trees, SNP, ...

- in large quantities \( (10^6 \) gene sequences)

- in large dimension \( (10^5 \sim 10^6 \) spots on DNA chips)
A possible solution: kernel methods

Kernel methods (partially) overcome these issues:

- Kernels for structured data
A possible solution: kernel methods

Kernel methods (partially) overcome these issues:

- Kernels for **structured data**
- **Operations on kernels** to integrate heterogeneous data
A possible solution: kernel methods

Kernel methods (partially) overcome these issues:

- Kernels for **structured data**
- **Operations on kernels** to integrate heterogeneous data
- **Regularisation** in order to deal with large dimensions
A possible solution: kernel methods

Kernel methods (partially) overcome these issues:

- Kernels for **structured data**
- **Operations on kernels** to integrate heterogeneous data
- **Regularisation** in order to deal with large dimensions
- **Statistical framework** for the processing of large datasets
What is a kernel?

- Let $\mathcal{X}$ be a set to be analyzed (e.g., gene sequences or protein structures).

- A kernel on $\mathcal{X}$ is a measure of similarity $K(x, x')$ between elements of $\mathcal{X}$ (that is symmetric and positive definite).

- Example: a kernel for finite-length sequences

\[ K(\text{aatcga}, \text{cgaagtagcccc}) = 0.4 \]
Geometric interpretation as inner product

If $K$ is a kernel on $\mathcal{X}$, then $\mathcal{X}$ can be mapped to a Hilbert space $\mathcal{H}$ through $\Phi : \mathcal{X} \rightarrow \mathcal{H}$ in such a way that:

$$\forall (x, x') \in \mathcal{X}^2, \quad K(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}.$$
Kernel trick

- Any algorithm for vectors that only involves inner products can be performed implicitly in the feature space by replacing the inner product by a kernel.
Kernel trick

- Any algorithm for vectors that only involves inner products can be performed implicitly in the feature space by replacing the inner product by a kernel.

- Example: Support Vector Machines (classification, regression), clustering, PCA, ICA, CCA, logistic regression... = kernel methods
Kernel trick

- Any algorithm for vectors that only involves inner products can be performed implicitly in the feature space by replacing the inner product by a kernel.

- Example: Support Vector Machines (classification, regression), clustering, PCA, ICA, CCA, logistic regression. = kernel methods

- “Simples kernels” can correspond to “complex” mappings.
Kernel trick

- Any algorithm for vectors that only involves inner products can be performed implicitly in the feature space by replacing the inner product by a kernel.

- Example: Support Vector Machines (classification, regression), clustering, PCA, ICA, CCA, logistic regression... = kernel methods

- “Simples kernels” can correspond to “complex” mappings

- Objects are not necessarily vectors!
Data representation with kernels

Each data set is a matrix

Kernel methods process these matrices
A few kernels for biological data

- Interpolated kernel for fixed-length sequences \((PSB'02)\)
A few kernels for biological data

- Interpolated kernel for fixed-length sequences \((PSB'02)\)
- Kernel for phylogenetic profiles \((ISMB'02)\)
A few kernels for biological data

- Interpolated kernel for fixed-length sequences (PSB’02)
- Kernel for phylogenetic profiles (ISMB’02)
- Kernel for molecular 2D structures (ICML’04)
A few kernels for biological data

- Interpolated kernel for fixed-length sequences (*PSB’02*)
- Kernel for phylogenetic profiles (*ISMB’02*)
- Kernel for molecular 2D structures de molécules (*ICML’04*)
- Mutual information kernel for sequences (*IJCNN’04*)
A few kernels for biological data

- Interpolated kernel for fixed-length sequences (*PSB’02*)
- Kernel for phylogenetic profiles (*ISMB’02*)
- Kernel for molecular 2D structures de molécules (*ICML’04*)
- Mutual information kernel for sequences (*IJCNN’04*)
- Local alignment kernel for sequences (*Bioinformatics 04*)
A few kernels for biological data

- Interpolated kernel for fixed-length sequences (*PSB’02*)
- Kernel for phylogenetic profiles (*ISMB’02*)
- Kernel for molecular 2D structures de molécules (*ICML’04*)
- Mutual information kernel for sequences (*IJCNN’04*)
- Local alignment kernel for sequences (*Bioinformatics 04*)
- Kernel for sets of points (*NIPS’04*)
Applications

- Signal peptide detection in protein sequences
Applications

- Signal peptide detection in protein sequences
- Gene function prediction
Applications

- Signal peptide detection in protein sequences
- Gene function prediction
- Virtual screening of small molecules
Applications

- Signal peptide detection in protein sequences
- Gene function prediction
- Virtual screening of small molecules
- Homology detection between gene sequences
Example 1: Phylogenetic profiles \( (ISMB\ 02) \)

<table>
<thead>
<tr>
<th>Gene</th>
<th>human</th>
<th>yeast</th>
<th>...</th>
<th>HIV</th>
<th>E. coli</th>
</tr>
</thead>
<tbody>
<tr>
<td>YAL001C</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>YAB002W</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

- Computed \textit{in silico}
- Useful to infer gene function
How to compare two profiles?

x
1 1 0 1 0 0 0 1 1 0

y
1 0 1 0 0 0 0 1 0 1
"Phylogenetic" kernel

\[ K(x, y) = \sum_e P(e)P(x|e)P(y|e), \]
### Gene function prediction (ROC 50)

<table>
<thead>
<tr>
<th>Functional class</th>
<th>Naive kernel</th>
<th>Tree kernel</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino-acid transporters</td>
<td>0.74</td>
<td>0.81</td>
<td>+ 9%</td>
</tr>
<tr>
<td>Fermentation</td>
<td>0.68</td>
<td>0.73</td>
<td>+ 7%</td>
</tr>
<tr>
<td>ABC transporters</td>
<td>0.64</td>
<td>0.87</td>
<td>+ 36%</td>
</tr>
<tr>
<td>C-compound transport</td>
<td>0.59</td>
<td>0.68</td>
<td>+ 15%</td>
</tr>
<tr>
<td>Amino-acid biosynthesis</td>
<td>0.37</td>
<td>0.46</td>
<td>+ 24%</td>
</tr>
<tr>
<td>Amino-acid metabolism</td>
<td>0.35</td>
<td>0.32</td>
<td>- 9%</td>
</tr>
<tr>
<td>Tricarboxylic-acid pathway</td>
<td>0.33</td>
<td>0.48</td>
<td>+ 45%</td>
</tr>
<tr>
<td>Transport Facilitation</td>
<td>0.33</td>
<td>0.28</td>
<td>- 15%</td>
</tr>
</tbody>
</table>
Example 2: Local alignment kernel (Bioinfo. 04)

- The Smith-Waterman local alignment score:
  \[ SW(x, y) = \max_{\pi \in \Pi(x,y)} s(x, y, \pi) \]

  is a widely-used measure of similarity between biological sequences, but... it is not a kernel

- The following local alignment kernel is valid:
  \[ K_{LA}^{(\beta)}(x, y) = \sum_{\pi \in \Pi(x,y)} \exp(\beta s(x, y, \pi)) , \]
Empirical evaluation

![Graph showing the number of families with given performance across ROC50 values for SVM-LA, SVM-pairwise, SVM-Mismatch, and SVM-Fisher.](image)
Comparison of heterogeneous data (NIPS’02)

VS

Detecting pathway activity? Data and graph denoising? Network reconstruction?
Using kernel-CCA
Example (ECCB’03)

Comparison of the metabolic network vs cell cycle gene expression in yeast
Correlated pathways
Extensions

- Feature extraction for gene supervised classification (*NIPS’02*)
- Feature extraction for gene clustering and operon detection (*ISMB’03*)
Supervised graph inference

Bayesian networks (Friedman et al., 2001), dynamical systems (Akutsu, 2000), nearest neighbour joining method (Marcotte et al., 1999)...
Graph learning through metric learning
Graph learning through metric learning

\[ \Phi \]
Graph learning through metric learning
Graph learning through metric learning
Graph learning through metric learning
Graph learning through metric learning

\[ \Phi \]
Graph learning through metric learning

\[ \Phi \]
Unsupervised graph learning

ROC curves: Direct approach
Supervised graph learning

ROC curves: Supervised approach

True positive vs False positive
Future of computational biology

- A strong and increasing demand to solve well-defined problems
Future of computational biology

- A strong and increasing demand to solve well-defined problems
- More and more possibilities to formulate new hypothesis/theories from results of data mining (e.g., scale-free properties...)
Future of computational biology

• A strong and increasing demand to solve well-defined problems

• More and more possibilities to formulate new hypothesis/theories from results of data mining (e.g., scale-free properties...)

• An urgent need for an adapted mathematical framework to represent and integrate biological data (probabilistic? kernel methods? dynamic systems? operator algebra?...)

Future of computational biology

- A strong and increasing demand to solve well-defined problems
- More and more possibilities to formulate new hypothesis/theories from results of data mining (e.g., scale-free properties...)
- An urgent need for an adapted mathematical framework to represent and integrate biological data (probabilistic? kernel methods? dynamic systems? operator algebra?...)
- How to transfer fundamental findings into applications, such as new therapies?
A challenge for the CAS-MPI Institute

- Seek a **fast international recognition** through an **original and high-level research**

- **Strong collaboration** with the CAS biological and medical facilities, and with the MPI excellence centers in computer science and mathematics

- Focus on a **small number of well-defined applications**, in collaboration with nearby CAS laboratories

- Keep a **long-term theoretical goal**
Acknowledgements

Collaborators at Kyoto University, University of Washington, UC Berkeley, UC Davis, MPI Tübingen, Institut Pasteur, Institut Curie, Paris 6 University