

## テーマ：構造データを扱う機械学習手法の研究

- 構造をもったデータを効率的に解析するための機械学習手法を研究しています
- 特にカーネル法とよばれる手法を用いた構造データの解析手法に貢献しました
- 現在は、生体ネットワークなどのネットワーク構造の解析手法の研究を行っています

生命システム情報学講座 バイオ情報ネットワーク分野  
指導教員：阿久津 達也 教授

鹿島 久嗣

## 構成

- 構造データの解析
  - ▶ 構造データとは
  - ▶ 構造データの解析はなぜ難しいか
- 解析手法
  - ▶ カーネル法による構造データの解析手法
    - カーネル法（畳み込みカーネル）とは
    - 構造データに対するカーネル関数の設計
      - [研究成果1] ラベル付順序木カーネルについて
      - [研究成果2] グラフカーネルについて
    - 構造データカーネルの発展
      - [研究成果3] 構造ラベル付けについて
  - ▶ ネットワーク構造をもったデータの解析手法
    - リンク予測について
    - [研究成果4] 遷移モデルに基づくリンク予測について
- まとめ／発表文献

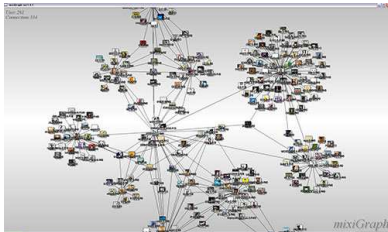
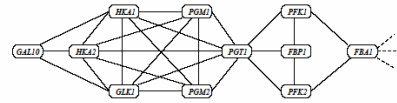
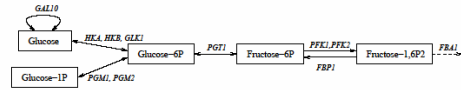
## 近年、構造をもったデータが増加しています

- 近年、構造をもった電子的データが増加している

- ▶ たとえば、

- 配列データ: DNA、タンパク質、自然言語、
    - 木構造データ: HTML/XML、RNA構造、構文解析木、系統樹、ディレクトリ
    - グラフ構造データ: 化合物、WWW、SNS、生体ネットワーク

- これらの解析が必要!



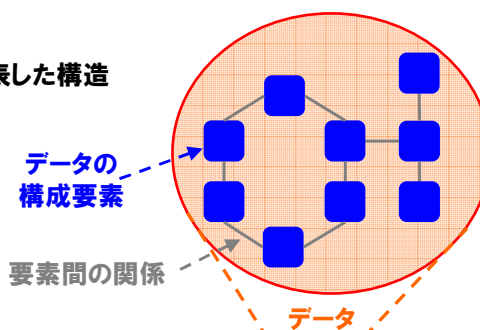
AGERTL	L	YVESIS	Y	I	VYLSGCT	F.	knowled
AGERTL	L	YVESIS	Y	I	VYLSGCT	F.	silivole
AGERTL	L	YVESIS	Y	I	VYLSGCT	F.	schason
AGERTL	L	YVESIS	Y	I	VYLSGCT	F.	shades
AGERTL	L	YVESIS	Y	I	VYLSGCT	F.	w/Africa
AGERTL	L	YVESIS	Y	I	VYLSGCT	F.	w/Thal-100
AGERTL	L	YVESIS	Y	I	VYLSGCT	F.	w/Thal-115
AGERTL	L	YVESIS	Y	I	VYLSGCT	F.	w/Korea
AGERTL	L	YVESIS	Y	I	VYLSGCT	F.	w/Vietnam
AGERTL	L	YVESIS	Y	I	VYLSGCT	F.	w/Salvador-1
AGERTL	L	YVESIS	Y	I	VYLSGCT	F.	w/Salvador-2
AGERTL	L	YVESIS	Y	I	VYLSGCT	F.	w/Brazil-1
AGERTL	L	YVESIS	Y	I	VYLSGCT	F.	w/Brazil-2
AGERTL	L	YVESIS	Y	I	VYLSGCT	F.	w/Honduras-1
AGERTL	L	YVESIS	Y	I	VYLSGCT	F.	w/Honduras-2
AGERTL	L	YVESIS	Y	I	VYLSGCT	F.	w/Panama

「構造データ」と一口に言っても...

「構造」には内部構造と外部構造の2種類があります

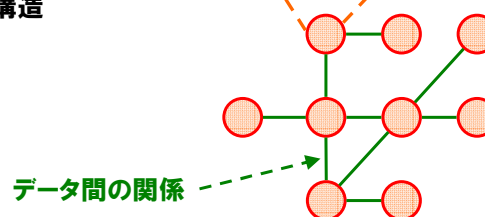
- 内部構造: データ内の要素の関連を表した構造

- ▶ HTML、XML
  - ▶ DNA
  - ▶ 化合物



- 外部構造: データ間の関連を表した構造

- ▶ Web
  - ▶ 社会ネットワーク
  - ▶ 遺伝子/蛋白質ネットワーク



私は、主に内部構造の取り扱いに取り組み、  
いくつかの研究成果を残しました

▪ 内部構造の扱い

▶ カーネル法に基づく解析手法

- [研究成果 1] ラベル付順序木カーネル
- [研究成果 2] グラフ・カーネル
- [研究成果 3] ラベル付けカーネルマシン

(▶ その他、パタンマイニングに基づくアプローチによる結果も)

▪ 外部構造の扱い（現在取り組んでいる）

▶ リンク予測問題への確率的アプローチ

- [研究成果 4] ネットワーク構造におけるリンク予測

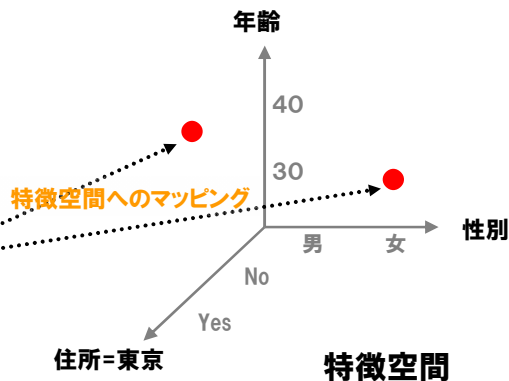
内部構造の扱いについての研究成果

従来の機械学習手法ではベクトル型のデータを前提としているため、  
構造をもったデータをうまく扱うことが出来ません

- 従来の機械学習手法では、データが特徴空間中のベクトルとして表されていることを前提とする
- 構造をもったデータでは、特徴空間の構成が自明でない

### 従来: ベクトル型のデータ

顧客番号	顧客氏名	年齢	性別	住所	...
0001	〇〇	40代	男性	東京都	...
0002	××	30代	女性	大阪府	...



構造を用いるための1つのアプローチ:  
「部分構造」を用いて特徴空間を構成する

- 構造データの解析では、「部分構造」が構造の性質を担っていると考える
  - ▶ 配列データの性質は、部分配列が担う (=マルコフモデル)
- たとえば、部分構造の有無や出現回数を使って特徴空間を構成し、ベクトル表現する

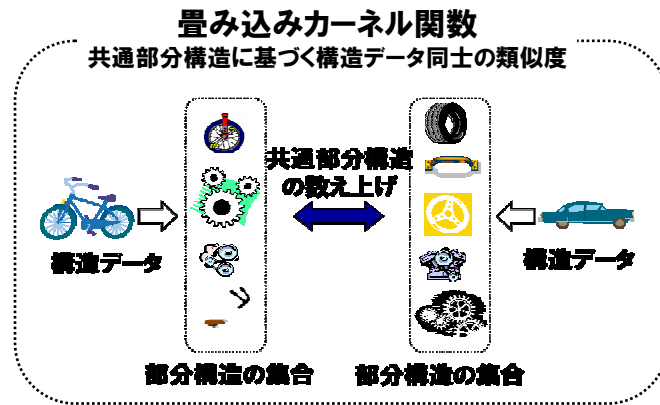
A C G C



配列番号	A-C	T-C	G-C	A-A	...
0001	1	0	1	0	...
...	...	...	...	...	...

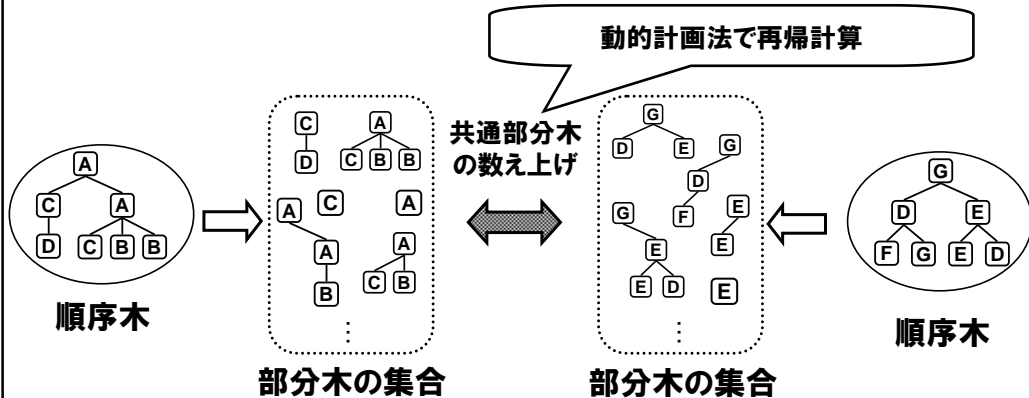
私は「畳み込みカーネル法」と呼ばれるすべての部分構造にもとづくアプローチに注目しました

- 畳み込みカーネル: 2つの構造データのカーネル関数(=2つのデータの類似度)を、共通にもつ部分構造を用いて定義する方法
  - ▶ 対象とする構造データに対して、部分構造の定義と、数え方の定義をする
  - ▶ カーネル関数の計算は、共通部分構造の数え上げの問題になる



**[研究成果1] 順序木に対するカーネル関数の設計を行いました**

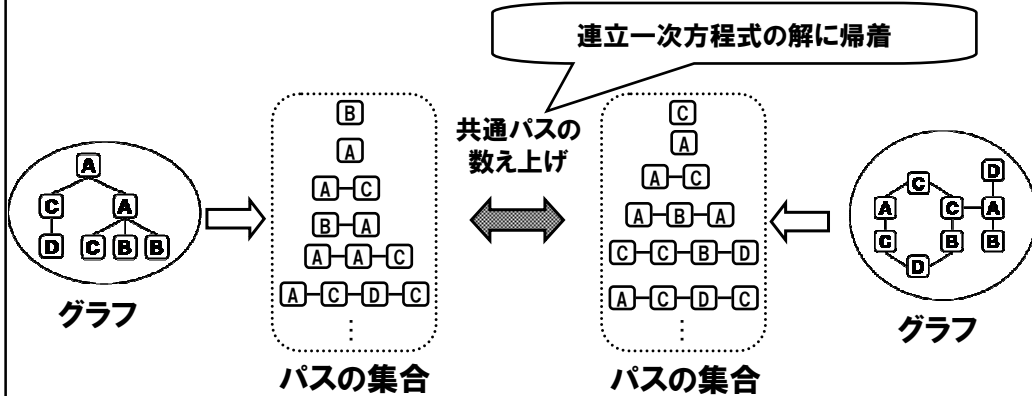
- 順序木カーネルでは、カーネル関数を
  - ▶ 部分構造=部分木
  - ▶ 数え方=2つの順序木に共通の部分木の個数によって定義する
- 難しいところ: 部分木は指数個ありうる  
 ← 解決法: 数え上げの計算を動的計画法によって再帰計算することで計算可能になる



H. Kashima and T. Koyanagi: Kernels for Semi-Structured Data, In Proc. 19th International Conference on Machine Learning (ICML), 2002

## [研究成果 2] グラフに対するカーネル関数の設計 を行いました

- グラフカーネルでは、カーネル関数を
  - ▶ 部分構造=パス
  - ▶ 数え方=2つのグラフからランダムウォークによって共通のパスが生成される個数によって定義する
- 難しいところ: パスは無限個ありうる
  - ← 解決法: 数え上げの計算を連立一次方程式に帰着することで計算可能になる

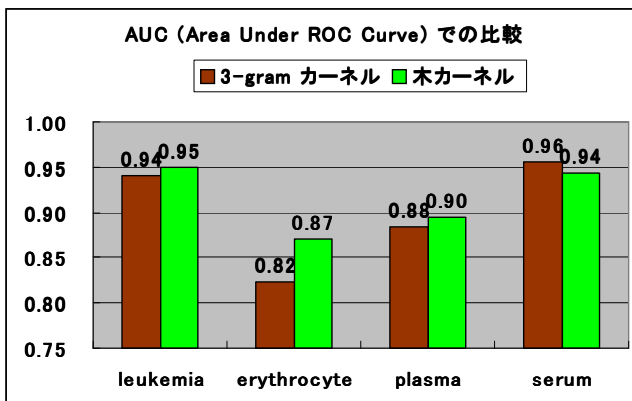
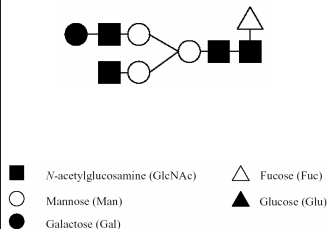


H. Kashima, K. Tsuda and A. Inokuchi: Marginalized Kernels Between Labeled Graphs, In Proc. 20th International Conference on Machine Learning (ICML), 2003.

## 応用: 木カーネルによる糖鎖の構造分類

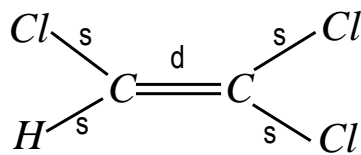
- 糖鎖の構造から組織を予測する問題に適用し、単純な方法 (3-gram に基づく方法) を若干上回る性能が得られた

### 糖鎖の木構造表現



## 応用：グラフカーネルによる化合物の毒性予測

- 指数時間の他手法（パタンマイニング）に匹敵する性能が得られた



化合物のグラフ表現

### パタンマイニング

MinSup	MM	FM	MR	FR
0.5%	60.1%	57.6%	61.3%	66.7%
1%	61.0%	61.0%	62.8%	63.2%
3%	58.3%	55.9%	60.2%	63.2%
5%	60.7%	55.6%	57.3%	63.0%
10%	58.9%	58.7%	57.8%	60.1%
20%	61.0%	55.3%	56.1%	61.3%

$p_v(v)$	MM	FM	MR	FR
0.1	62.8%	61.6%	58.4%	66.1%
0.2	63.4%	63.4%	54.9%	64.1%
0.3	63.1%	62.5%	54.1%	63.2%
0.4	62.8%	61.9%	54.4%	65.8%
0.5	64.0%	61.3%	56.1%	64.4%
0.6	64.3%	61.9%	56.1%	63.0%
0.7	64.0%	61.3%	56.7%	62.1%
0.8	62.2%	61.0%	57.0%	62.4%
0.9	62.2%	59.3%	57.0%	62.1%

### グラフカーネル

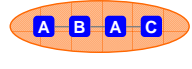
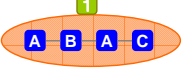
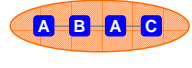
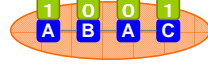
## これらの研究はその後、さまざまな発展を遂げています

- 順序木カーネル
  - ▶ 曖昧マッチングのとりこみ [久保山ら, 2006]
- グラフカーネル
  - ▶ ランダムウォークの設計による高精度化 [Mahe et al., 2004]
  - ▶ Cyclic Pattern [Horvath et al., 2004]、Shortest Path [Borgwardt & Kriegel., 2005] による高速化
  - ▶ タンパク質立体構造分類への適用 [Borgwardt et al., 2005]
- より複雑な問題への適用
  - ▶ 構造データのラベル付け問題への適用 [Kashima et al., 2004]

その後の発展の一例...

**[研究成果 3] 構造データのラベル付け問題への適用 を行いました**

- ラベル付け問題は、構造データの各要素に対してラベルを割り当てる予測問題
- ラベル付け問題の例：
  - 形態素解析、固有表現抽出、タンパク質の2次構造予測、遺伝子領域の予測、...
- 各要素の分類問題と考えることでカーネル法を適用できる

	入力	出力
分類問題	 構造データ	 構造データのクラス
ラベル付け問題	 構造データ	 各要素のラベル

H. Kashima and Y. Tsuboi: Kernel-Based Discriminative Learning Algorithms for Labeling Sequences, Trees and Graphs, In Proc. 21st International Conference on Machine Learning (ICML), 2004.

**応用：自然言語文からの情報抽出**

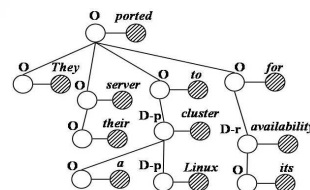
- 局所的な情報 (bi-gramモデル) に基づく方法を上回る性能が得られた
  - 固有表現抽出
    - 人名、組織名、場所名などを示すフレーズの抽出
  - 製品使用情報抽出
    - ある製品を使っているという事実を抽出
      - {製品名/企業名/数量/理由} × {導入/導入検討中/導入しない} など
    - 構文解析木構造も用いることで精度を向上できる

表 1 固有表現抽出結果 (括弧内は標準偏差)

	精度	適合率	再現率	F1
配列カーネル	88.7% (3.4)	49.0% (6.0)	23.1% (8.1)	30.5 (6.7)
HM パーセプトロン	80.2% (11.5)	23.8% (14.6)	17.9% (3.0)	18.6 (5.2)

表 2 製品使用情報抽出結果 (括弧内は標準偏差)

	精度	適合率	再現率	F1
SEQUENCE KERNEL	89.7% (2.0)	52.2% (9.5)	29.6% (4.0)	37.5(4.1)
TREE KERNEL	89.9% (2.7)	51.4% (10.9)	32.5% (14.4)	38.9 (12.1)
HM-PERCEPTRON	89.7%(1.8)	51.5%(8.5)	24.0%(21.4)	28.9(20.3)





外部構造の扱いについての研究成果

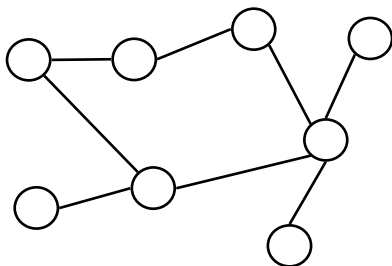
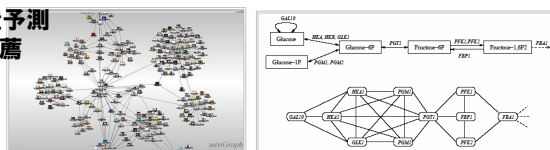
外部構造の解析としては、リンク予測の問題に取り組んでいます

▪ リンク予測問題：部分的に観測されているネットワーク構造から、残りの構造を推定する問題

- 例：
  - 生体ネットワークの構造予測
  - SNSにおける友人の推薦

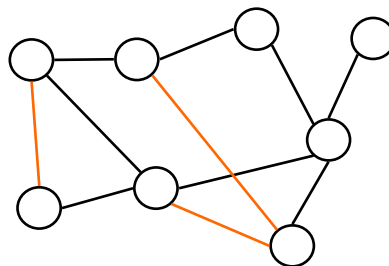
▪ リンク予測に用いることのできる情報

- ▶ 各データ (= ノード) のもつ情報
- ▶ 関係 (= 構造) のもつ情報



部分的に観測される構造

補完  
⇒



予測される構造

**[研究成果 4] ネットワーク構造の時間変化モデルの学習によって、ネットワークの構造情報のみからリンク予測を行う手法を提案しました**

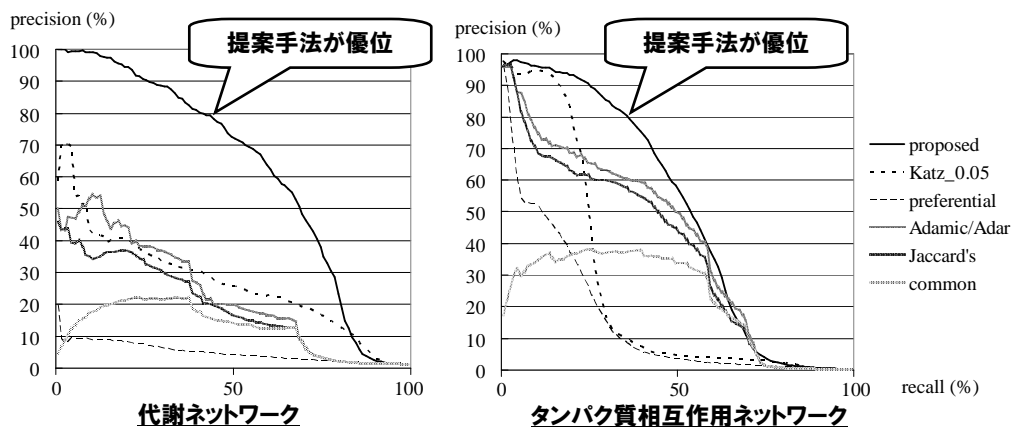
- ネットワーク構造の時間変化モデルを考える
  - ▷ ノードからノードへの枝の「コピー & ペースト」が確率的におこるモデルを仮定する
    - “node copy model”のパラメトリック版
- 観測されるネットワークは、この時間変化モデルの定常状態であると仮定して、パラメータを学習、残りの部分を予測する



(投稿中) H. Kashima and N. Abe: A Parameterized Probabilistic Model of Network Evolution for Supervised Link Prediction, submitted to *IEEE International Conference on Data Mining (ICDM)*, 2006.

**応用: 生体ネットワーク構造予測**

- 提案されているいくつかのリンク予測指標による予測を上回る性能が得られた



## まとめ：構造データを扱う機械学習手法の研究

- 構造をもったデータを効率的に解析するための手法を研究しています
- 特にカーネル法とよばれる手法を用いた、内部構造をもつデータの解析手法に大きく貢献しました  
(▶ その他、パタンマイニングに基づくアプローチにも貢献\*)
- また、最近では外部構造をもつデータの解析手法の研究を行っており、研究成果が出つつあります
  - ▶ 現在はさまざまなネットワーク生成モデルに基づくリンク予測モデルを開発中

\* A. Inokuchi and H. Kashima: Mining Significant Pairs of Patterns from Graph Structures with Class Labels, In Proc. 3rd IEEE International Conference on Data Mining (ICDM), 2003.

## これまでの研究成果：主な査読付発表論文等（赤字は本発表に関連するもの）

- ジャーナル論文
  - ▶ 鹿島, 坂本, 小柳: 木構造データに対するカーネル関数の設計と解析, 人工知能学会論文誌, Vol.21, No.1, 2006.
  - ▶ 鹿島, 津村, 井手, 野ヶ山, 平出, 江藤, 福田: ネットワークデータを用いた分散システムにおける異常検出, 電子情報通信学会論文誌, Vol. J89-D, No. 2, 2006.
  - ▶ T. Shibuya, H. Kashima and A. Konagaya: Efficient Filtering Methods for Clustering cDNAs with Spliced Sequence Alignment, *Bioinformatics*, 2004.
  - ▶ T. Fukao, H. Kashima and N. Adachi: Decentralized Adaptive Control with Improved Transient Performance, 計測自動制御学会論文誌, Vol.35, No.7, pp. 869-878, 1999.
- 主な国際会議論文
  - ▶ (投稿中) H. Kashima and N. Abe: A Parameterized Probabilistic Model of Network Evolution for Supervised Link Prediction, submitted to *IEEE International Conference on Data Mining (ICDM)*, 2006.
  - ▶ H. Kashima: Risk-Sensitive Learning via Expected Shortfall Minimization (Extended Abstract), In Proc. *2006 SIAM Conference on Data Mining (SDM)*, 2006.
  - ▶ H. Kashima, T. Tsumura, T. Ide, T. Nogayama, R. Hirade, H. Etoh and T. Fukuda: Network-Based Problem Detection for Distributed Systems, In Proc. *21st International Conference on Data Engineering (ICDE)*, 2005.
  - ▶ T. Ide and H. Kashima: Eigenspace-based Anomaly Detection in Computer Systems, In Proc. *10th ACM SIGKDD Conference (KDD)*, 2004.
  - ▶ H. Kashima and Y. Tsuboi: Kernel-Based Discriminative Learning Algorithms for Labeling Sequences, Trees and Graphs, In Proc. *21st International Conference on Machine Learning (ICML)*, 2004.
  - ▶ A. Inokuchi and H. Kashima: Mining Significant Pairs of Patterns from Graph Structures with Class Labels, In Proc. *3rd IEEE International Conference on Data Mining (ICDM)*, 2003.
  - ▶ H. Kashima, K. Tsuda and A. Inokuchi: Marginalized Kernels Between Labeled Graphs, In Proc. *20th International Conference on Machine Learning (ICML)*, 2003.
  - ▶ H. Kashima and T. Koyanagi: Kernels for Semi-Structured Data, In Proc. *19th International Conference on Machine Learning (ICML)*, 2002.

## これまでの研究成果：その他（赤字は本発表に関連するもの）

### ▪ 本／解説記事

- ▶ Hisashi Kashima, Koji Tsuda and Akihiro Inokuchi: Kernels for Graphs, in Kernel Methods in Computational Biology, MIT Press, 2004.
- ▶ 鹿島 久嗣: カーネル法による構造データマイニング, 情報処理, Vol. 46, No. 1, 2005.
- ▶ 鈴木 英之進, 鹿島 久嗣 (編): 特集「最新！データマイニング手法」, 情報処理, Vol. 46, No. 1, 2005.

### ▪ チュートリアル講演等

- ▶ 鹿島 久嗣: 「ネットワーク構造解析 - 機械学習によるアプローチ」, 人工知能学会 第63回人工知能基本問題研究会 (SIG-FPAI), 2006/9/8
- ▶ 鹿島 久嗣, 坪井 祐太, 工藤 拓: 「言語処理における識別モデルの発展 -- HMMからCRFまで --」, 言語処理学会第12回年次大会, 2006/3/13
- ▶ 鹿島 久嗣: 「構造データマイニングの手法とバイオインフォマティクスへの応用」, 化学工学会 第37回秋期大会, 2005/9/17.
- ▶ 鹿島 久嗣: 「カーネル法による構造データの解析」, 電子情報通信学会 パターン認識・メディア理解研究会, 2005/2/25.

## まとめ

- 構造をもったデータを効率的に解析するための手法を研究しています
- 特にカーネル法とよばれる手法を用いた、内部構造をもつデータの解析手法に大きく貢献しました  
(▶ その他、パタンマイニングに基づくアプローチにも貢献\*)
- また、最近では外部構造をもつデータの解析手法の研究を行っており、研究成果が出つつあります
  - ▶ 現在はさまざまなネットワーク生成モデルに基づくリンク予測モデルを開発中

\* A. Inokuchi and H. Kashima: Mining Significant Pairs of Patterns from Graph Structures with Class Labels, In Proc. 3rd IEEE International Conference on Data Mining (ICDM), 2003.