ISSN 1349-516X

NII Today

National Institute of Informatics News

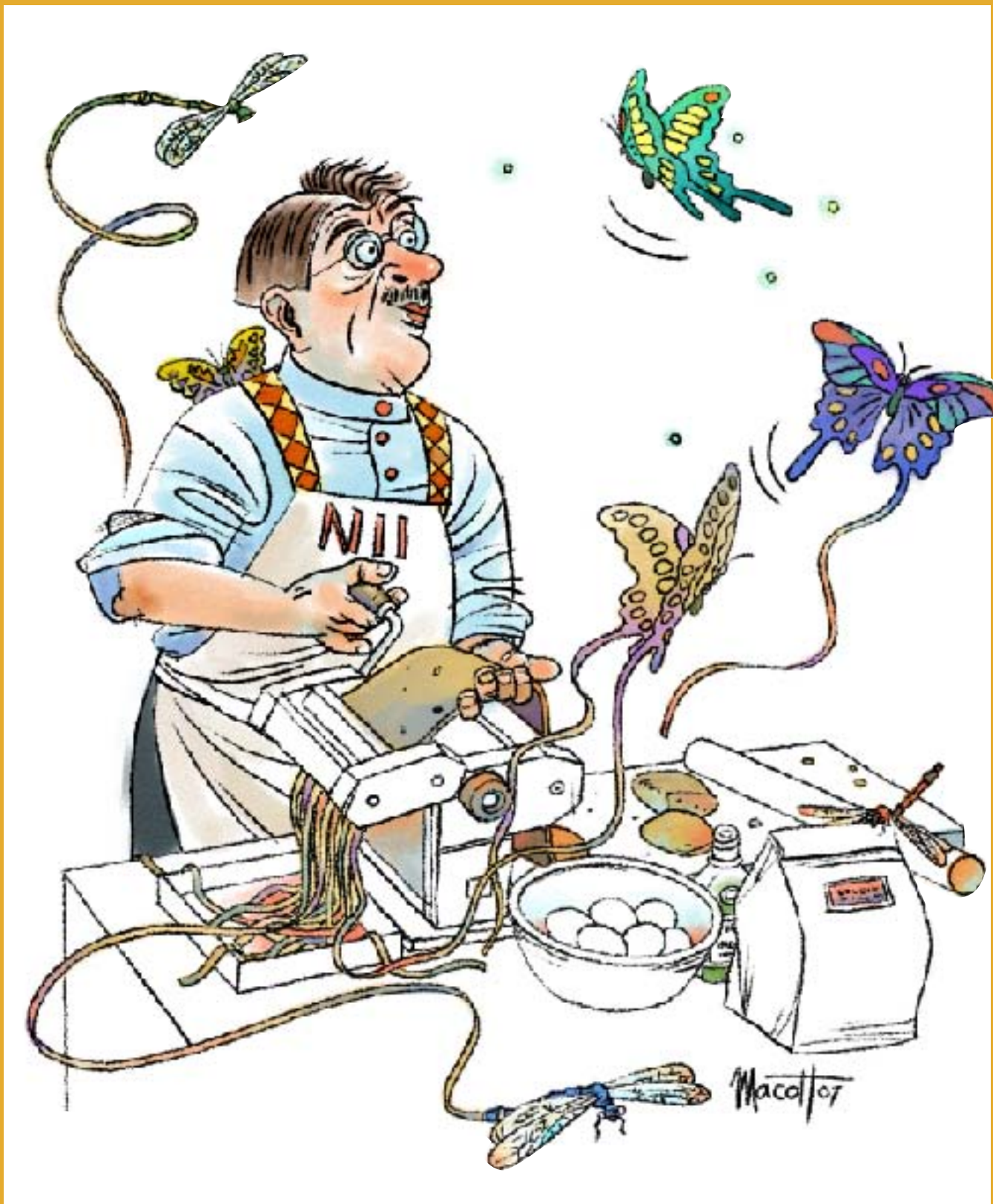
国立情報学研究所ニュース
No.37 September 2007

特集 融合の情報学

ゲノムは融合の要
効率的なアルゴリズムで分子構造の解明が速く、正確に

「情報リンケージ」が生み出す新たな価値

NIIが誇る大学院教育



ゲノムは融合の要
―一体化する生命科学と情報学

〔背景〕21世紀はじめにヒトゲノム計画が一通り終了し、生命科学は情報爆発の時代を迎えている。特にDNAやタンパク質などに関するデータの増大は著しく、それらの相互作用や関係を記述するデータベース、研究成果を報告する論文も増える一方だ。大量の情報や知識から、必要とするものを取り出せるようにするために、専門分野によってばらつきのある用語や概念を統制し、知識を体系化して、コンピュータで記憶・計算できるように整える努力が進んでいる。こうした努力は、研究の発展に役立つばかりでなく、生命科学の知識を広めることにも貢献するはずだ。生命科学と情報学は、分かち難い間柄になっている。

藤山秋佐夫

ふじやま・あさお
国立情報学研究所　教授
総合研究大学院大学　教授

南　生命科学と情報学の融合という切り口から、ご専門のゲノム科学についてうかがいたいと思います。大きなステップになったヒトゲノム計画も、コンピュータなしには達成できませんでしたね。

藤山　ヒトゲノム計画が具体的にスタートした1989年頃には、すでにDNAについていくつかのデータベースがありました。日本のDDBJ（DNA Data Bank of Japan）、米国のGen Bank、欧州のEMBL（European Molecular Biology Laboratory）が代表的なものです。予想されるヒトゲノムのデータ量からみて、コンピュータが必要になることは当初から明らかでしたが、どの程度必要かはまったく想像を超えていました。

南　当時と比べると、コンピュータや通信など、研究を支える情報環境も大きく変化しましたね。

藤山　80年代には大容量データは磁気テープにして航空便でやり取りしていたものです。通信の高速化とネットワーク化は、ユーザーとして研究を進めるうえで大いにプラスです。

南　情報学の専門家と生物学の専門家が緊密に協力するようになったのはいつ頃からですか。

藤山　80年代はじめ頃は、生物学者も必要に応じて自分でプログラムを書いていましたが、情報学の専門家から見たら子供みたいなもの。だんだんプロに任せるようになったのです。しかし、生物学者と情報学者がいきなり顔を合わせても話が通じません。互いの言葉を共有する努力をして、なんとか意思疎通をはかるのに精一杯でした。生物学と情報学が結婚できるか？　という議論までありました。

南　今はシナリオ通りの円滑な関係になっていますか。

藤山　現在は、いくつかの大学で、はじめから両方の訓練をする場ができています。ゲノムを研究すると、生物全体についての情報を扱うことになり、それぞれの関係を見たり、異なる生物のゲノムを比較したりしますから、コンピュータを使わざるを得なくなります。しだいにこの分野の情報学で仕事をしていける人たちが育ってきます。また、大学だけではなく、製薬企業などでも、ゲノムや遺伝情報を企業レベルで活用する必要性が生まれ、生物情報学の人材が求められるようになる。そういうかたちで融合分野が根付いていくでしょう。

しかし、融合は、はじめから簡単に受け入れられるものではありません。本当に新しいものは、誰もその価値を評価できないから、世の中に簡単には受け入れられないものでしょう。

南　ゲノム科学もそうでしたか。

藤山　はじめは「あんなものは科学ではない」と言われていました。今ではゲノム情報を誰でも使うようになり、使わないと論文が書けません。

分子生物学だって、完全に定着するのに20～30年はかかったでしょう。私が学生の頃、60年代後半～70年代前半の初期の分子生物学者たちは、中央には受け入れられず、地方の大学で苦労されていました。当時、分子生物学は、化学、物理学、生物学の融合分野だったわけです。当事者たちは特にそんな意識はなく、新しい学問をつくる意気込みでやっていたのですが……。今では、分子生物学会は会員数8000人規模になり、基礎生物学最大の学会です。

南　情報は融合を進める手段でもありますね。

藤山　その通りです。ゲノムのデータは公開されていて、誰でもインターネットを通じて無料でアクセスできるので、生命科学の専門家はもちろん、ほかの分野の人たちが研究するにも重要な役割を果たしています。

ゲノム研究者のなかには、医学者、植物学者、動物学者など多様な背景をもつ人がおり、さらに、工学、農学、薬学など広い応用分野の研究者がゲノム情報を土台にして仕事をしています。専門が違うと、同じ遺伝子や酵素でも呼び方が違うなど、うまく言葉が通じないことがあります。さまざまのデータベースを統合したり、ゲノムを比較したりするにも、用語の違いが壁になりかねません。

Jabion
日本語バイオポータルサイト

最新のバイオ情報を
分かりやすい日本語で

そこで、今、使おうとしているのが、自然言語処理やオントロジーなどの情報理論です。用語や概念の体系化をして、どんな分野の人でも使える辞書を作ろうとしています。大量のデータが生産されていますから、求める情報や知識をうまく抽出する技術も作り上げる必要があります。

南　情報には、この分野を社会に伝達する役割もあるのではないですか。

藤山　情報を発信せよという政策的な力は強く働くようになりましたが、研究している当人たちにはその余裕がないのが実状です。

私たちは「日本語バイオポータル」というサイトを作っています。はじめた理由のひとつは、新聞の科学記事は一般の人々には少し難しいのではないかと思うからです。紙面の制限がありますから、背景が十分に伝わりにくい場合もあります。社会的に注目を集めるような生命科学のテーマを、できるだけわかりやすく提供したいと考えています。

インタビュアーの一言

NIIに生物学者がいて、生物学と情報学がこんなに融合しているとは！　ゲノム研究のはじまりから、ずっとこの分野の先駆けである藤山先生は、実験を続ける一方、インターネットで今の生物学を一般の人たちに知ってもらおうとプロジェクトを立ち上げた。新聞では、基礎的な科学についての記事はまだきわめて少なく、一般記事のわずか0.1%というデータもあるそうだ。研究者が自ら発信するポータルサイトはたいへん貴重だと思う。ご活躍に期待しながら、注目したい。

南砂

みなみ・まさご
読売新聞東京本社　編集委員

NII Special

融合の情報学

CREATING FUSION FROM TRANSDISCIPLINARY

効率的なアルゴリズムで分子構造の解明が速く、正確に

Hiroko Satoh
佐藤寛子
情報学プリンシプル研究系准教授

Takeaki Uno
宇野毅明
情報学プリンシプル研究系准教授

異なる分野間の研究の融合は、個々の分野では生み出しえない成果をもたらす。ただし、学術的に優れた成果を得るためには、お互いに相手の研究を理解し、得意とする専門を研究に有効な形で生かす必要がある。こうした「本物の融合」が、NIIを舞台に、化学と情報学の間で進んでいる。

世の中には多くの有機化合物がある。天然に存在するものもあれば、人間が作り出したものもある。そうした化合物を薬や材料などに利用する場合、その分子の形（分子構造）を知ることは重要だ。このため、化学者はさまざまな手法を駆使して分子構造を解明する。なかでも核磁気共鳴（NMR）法は強力だ。とくに炭素13のNMR（^{13}C-NMR）からは、分子の骨組み（炭素の骨格）がわかるので、有機化合物の構造を知るのに欠かせない。このデータの1つに「化学シフト」という重要な値がある。これは、炭素原子の周りの構造に大きな影響を受けるので、分子構造を知る大きな手がかりとなる。

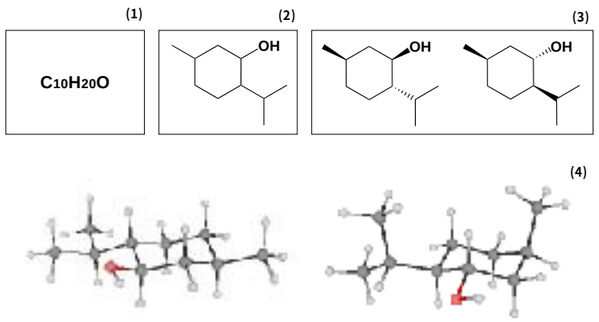


図1：メントール（ハッカの香り）の分子構造のさまざまなフェーズ。（1）分子式。分子を構成する原子の種類と個数で表したもの。（2）構造式（平面）。原子と結合関係だけを表したもの。（3）構造式（立体化学）。（2）に結合の向きの情報が付されたもの。太線は紙面よりこちら側、点線は紙面より向こう側の向きの結合を意味する。（4）3次元構造。（3）と（4）は左から l - メントール、d - イソメントール。香りが少し異なる。

これまでに、多くの有機化合物について、^{13}C-NMRの化学シフトデータが得られている。これらは、正しく効率的に使えば、化学シフトや構造の予測に威力を発揮する。そのためには、分子情報を化学的に意味のある形で高度に統合し、そこから適切に検索できることが必要だ。これはまさに、化学と情報学の融合を必要とするシステムである。NIIの佐藤寛子准教授は、さまざまな分野の研究者と協力してそのようなシステム — CAST/CNMRを構築した。

分子の立体的な構造をコンピュータに入れよう

一般に分子構造データは原子の位置座標と結合情報で表される。これはグラフィックス表示には便利だが、検索や予測には不向きだ。検索や予測のためには、分子を特徴によって分類・整理する必要がある。佐藤准教授は、理化学研究所（理研）の基礎科学特別研究員だった1996年、このテーマに取り組み始めた。もともと化学の出身だが、膨大な化学情報を化学現象の解明に活かすことを大きな目標としていた。

一言で「分子構造」といっても、実は奥が深い。原子の種類と数で表される分子式から立体構造まで、さまざまなフェーズがある（図1参照）。図1(3),(4)にあるように、分子は本来、立体的な形をしているため、平面構造は同じでも、実は立体構造が違い、異なる分子だということが起こりうる。しかも、このような違いが分子の性質や反応に大きく影響する。同じ構造でも違った描き方ができるので、コンピュータに入れる場合は、同じものは同じ表現にまとめた上で検索しやすい形に加工する。これがコード化だ。

佐藤准教授は、反応や物質の予測に不可欠な立体構造もコンピュータに入れようと考えていた。一方、分子構造解析の研究者である越野広雪・理研先端技術開発支援センター物質構造解析チームチームリーダー（当時は分子構造解析室研究員）は、日々の構造解析に活用できるデータベースを探していたが、どれも立体構造を満足に扱えないという致命的な問題があった。当時、立体的な情報を扱うコードはいくつかあったが、厳密に記述できるものはなかった。雑談での情報交換をきっかけとして目的が一致した2人は、「立体構造のコード化と化学シフト予測」の研究を共同で開始した。

この研究により、まず、CAST（CAnonical-representation of STereochemistry）というコード化の方法が開発された。これは、平面構造をコード化したものに、立体構造の情報と後述する環構造の情報を加えたものだ。「分子は、平面構造も立体構造もそれぞれ特徴をもっているのですが、このすべてを取り込むコンピュータ表現を考案するのは大変な作業でした」と佐藤准教授は振り返る。試行錯誤の結果、立体構造の表現に汎用的でシンプルなやり方を採用した。結合の二面角（3本の結合が連続しているときに、前の2本が作る平面と、後の2本が作る平面のなす角）を利用する方法である。

アルゴリズムのプロが仲間に加わった

佐藤准教授がソフトウエアを開発する一方、越野チームリーダーは文献や自身の実測データの中からデータベースに収めるのにふさわしい信頼できるものを選び出し、検証を重ねながらシステムを構築していった。こうして、2000年頃までには、「注目する炭素原子を含む部分構造をCASTコード化→これをクエリ（検索語）としてデータベースを検索→ヒットした化合物の化学シフトから注目炭素の化学シフトを予測」という手続きを行うCAST/CNMRシステムが一通り完成した。

「しかし、まずは動くものを作ることを優先させたので、コード化も検索も遅いという難点がありました。また、初期のCASTコードには、対称性の高い分子のコードが1つにまとまらないなどの細かなゆらぎがありました」。佐藤准教授は、NII内の情報交換を目的とした昼食会で、この問題をとりあげた。これを聞いた宇野毅明准教授は、「大学受験のとき、数学の次に化学が好きだった自分なら、佐藤先生の研究に絡めそうだ」と感じたという。宇野准教授の専門は、アルゴリズムの開発。ハードウエアを変えずにプログラムの設計を変えて計算を速くする方法を考える。宇野准教授は、アルゴリズム研究の仲間である岩田覚（さとる）・京都大学数理解析研究所准教授（当時は東京大学大学院助教授）にも声をかけ、2003年から3人の共同研究が始まった。

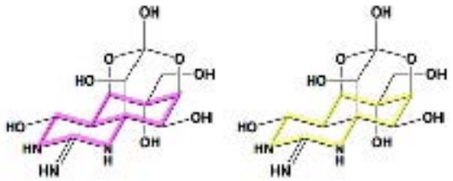


図2: テトロドトキシン（フグ毒）の構造。黄色の環は、ピンクの環の組み合わせで作ることができる。複数の環の組み合わせでできる環を除外すると8種類、除外しないと12種類の環がある。

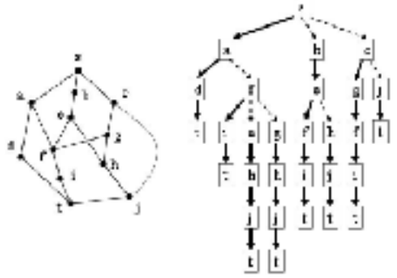


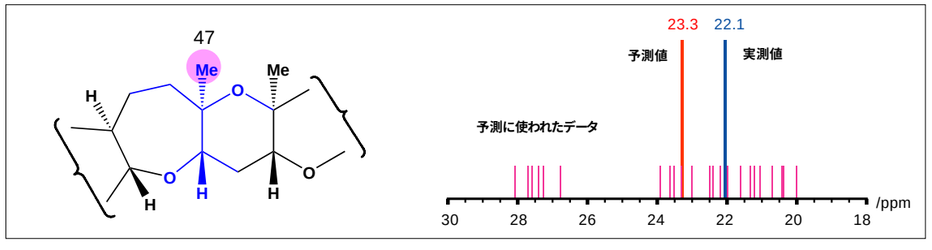
図3：環構造列挙アルゴリズムの行う探索の例。左図のsからtへの経路を列挙して環構造を見つける（sとtを結ぶ線は省略してある）。アルゴリズムはsから出発して、隣り合うそれぞれの点へ移動して探索するが、そのようすを示したのが右図。頂上のsから末端のtまでの経路それぞれが、左図での経路に対応する。

宇野准教授と岩田准教授は、問題点をすぐに見抜いた。平面構造をコード化するときは、分子の中のある原子を出発点とし、そこから１本の結合を介したところにある原子、2本の結合を介したところにある原子・・・と、順番に取り上げていく。しかし、同じ数の結合を介した原子がいくつもあるとき、どれを先に取り上げるかのルールを厳密に決めておかないと、同じ分子に別々のCASTコードがつけられる可能性がある。

アルゴリズムの研究者にとって、この種の問題は得意分野である。「取り上げる順番が何通りもあるのなら、逆に、すべての可能性を試してしまうのです。コードは文字のつながりと思っていいですから、出てきたすべてのコードの中から、あいうえお順で最初に来るものを採用すればいい。このやり方なら、誰がやっても、同じ平面構造には同じコードが割り当てられます」（宇野准教授）。原子や結合の種類などで優先順位をつけようという化学者の発想とは、根本的に違う情報科学者の発想だ。これで検索の精度が向上した。

さらに、コード化の計算時間も短縮された。以前は、コードを割り当てる際に再計算を繰り返していたため、非常に時間がかかった。そこで、アルゴリズム理論から高速化の方法を探り、再計算しないですむロジックを発見した。これによりコード化の速度は2桁以上速くなり、原子数が400個ほどの分子でも、1～２分でコード化できるようになった。

青色で表した部分の平面構造と立体構造が似ている分子を検索し、その化学シフトを平均したものを予測値とした場合



青色で表した部分の平面構造と立体構造が似ているだけでなく、黄色で示した環構造をもつ分子を検索し、その化学シフトを平均したものを予測値とした場合

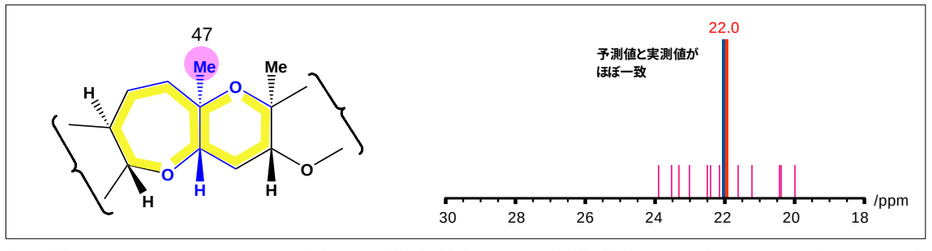
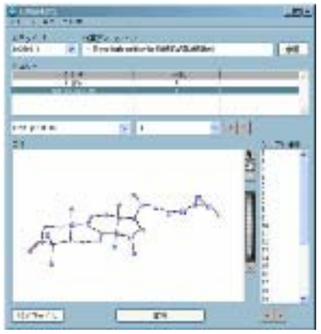
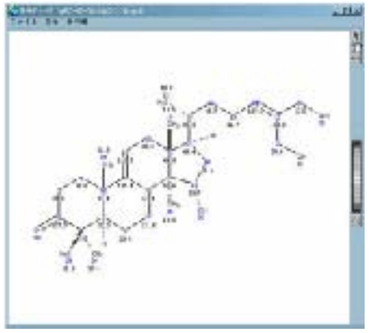


図4：環構造を入れたCASTコードにより、NMR化学シフト予測精度が向上した例。海洋生物が生産する毒素プレベトキシンBの分子のうち、47番目の炭素原子（ピンクの丸で示したメチル基に含まれている）の化学シフトを予測した。



予測条件を入力する画面



予測された^{13}C化学シフト値

図5: CAST/CNMRシステムの実行画面。化学シフトを予測したい分子の構造を与え、どういう条件で予測するかを決めると（左）、4000種ほどの化合物が収められたデータベースを検索し、それぞれの炭素について化学シフトが予測される（右）。炭素原子を選ぶと、その予測に使われた分子のデータも見ることができる。

理想的な融合の形

さて、前述の通り、化学シフトの値は炭素原子の周りの構造に大きく影響を受けるが、分子によっては環構造も大きく効いてくる。CASTコードに環構造の情報も加えてあるのはそのためだ。

しかし、分子がどのような環を含んでいるかを、コンピュータに列挙させることは意外に難しい。例えば、図2の分子の中の12個の環構造をもれなく重複なく列挙するには、数学的に裏付けられた計算方法が必要となる。情報科学の分野では、環構造列挙のアルゴリズムは1970年代に開発されていた。しかし、化学シフト予測に利用する場合、予測に必要な環構造情報のみを正確に、かつ効率的に取り出す必要があった。そのために、宇野准教授が次に取り組んだのは「複数の環の組み合わせではない環だけを列挙するアルゴリズム」だった。

「2003年頃、佐藤先生と茶飲み話の中で相談されました。過去のアルゴリズムでは複数の環の組み合わせでできる環もすべて見つけるので計算に時間がかかる。しかし、1つロジックを組み込むことで、比較的簡単に改良できました」（図3）と、宇野准教授は振り返る。これにより、複雑な環構造をもつ分子でも、ほんの一瞬で計算できるようになった。

「これらのアルゴリズムを組み込み、CAST/CNMRシステムの実用性は大きく前進しました」と佐藤准教授は言う。適切な環の情報を取り込むことで化学シフトの予測精度はさらに向上した（図4）。また、「このシステムの予測と実際の化学シフトの値が合わないことがきっかけとなって、過去の構造決定の間違いを訂正した成果もいくつか出てきています」（越野チームリーダー）。

GUI（グラフィカルユーザインターフェース）を備えたテスト版（図5）も完成しており、越野チームリーダーが検証と評価を進めている。佐藤准教授は、「構造解析の研究者を中心に広く利用してもらえるようにしたい」と今後のCAST/CNMRシステムの展開を語る。

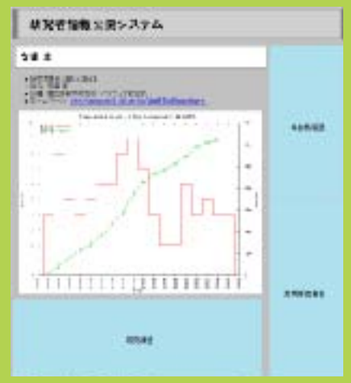
今回の共同研究について宇野准教授は、「自分たちアルゴリズムの研究者だけで研究をすると、解こうとする問題は理論家が考えた想定の上にとどまってしまいます。しかし、現実の問題は想定した問題と違い、面白いひねりが入っています。そういう問題を提示されて解法を見つけると、実際に役に立つことがうれしいし、研究にも深みが出てきます」と十分な手応えを感じたようだ。

一方、「化学データは複雑で、研究を進めていくと必ず多様な問題にぶつかります。これを根気よく1つずつ解決しないと、化学的に価値のある成果は得られません。宇野先生と岩田先生、プログラムを開発した東京大学大学院生の小市俊悟さんが、こうした複雑性を認識して粘り強く研究を共に進めてくれたことが、大きかったと思います」と佐藤准教授。情報学と化学の理想的な融合は、これからも着実に成果を上げていくことだろう。

（取材・構成　青山聖子）

「情報リンケージ」が生み出す新たな価値

同一の人やものに関する情報を集めて一元管理する「情報リンケージ」を利用して、情報に新たな価値を見出そうとする研究者たち。「情報リンケージ」というテーマで集まった彼らは、どのように価値ある成果を生み出そうとしているのだろうか。



Google Maps

さまざまなソースからの情報を集めて作成した「研究者情報公開システム」の例。一覧表、グラフ、図などで表示される。左のグラフは参加プロジェクト数（赤の棒グラフ）と成果リスト登録件数（緑の折れ線グラフ）。右の地図は共同研究者（上位100名まで）の分布。安達　淳教授は、NII学術基盤推進部の部長で、新領域融合プロジェクトのメンバーである。

Akiko Aizawa
相澤彰子
コンテンツ科学研究系教授

一般には「名寄せ」という名で知られる「情報リンケージ」(*1)。「目立ちませんが、じつは非常に重要な操作です」と語るのは、国立情報学研究所新領域融合プロジェクト(*2)で「大規模リンケージ情報の収集・分析手法」グループの代表を務める相澤彰子教授。もともと「情報リンケージ」を用いることになったきっかけは、「論文のデータベースや図書のオンラインカタログなどの重複登録をなくすこと（クリーニング）でした」という。

たとえば論文の著者に同じ名前があった場合、それが重複なのか同姓同名の別人なのかは、データベースを利用する側にとって重大な問題である。情報リンケージによってデータベースの質を高めることは、データを使用するときに大きなメリットとなる(次ページ図)。

コンピューターの普及とネットワークの巨大化により情報の量は増え続け、しかも手に負えないほど乱雑にばらまかれている。情報を必要に応じて整理整頓する情報リンケージが必要とされる場面は、それこそ無数にある。

情報リンケージとデータベース

一口に情報リンケージといってもさまざまなパターンが考えられる。相澤教授は「難易度から問題設定は4段階に分けられます」という。

まずは先の例で示した、1つのデータベース内での操作である。データベースは、ある目的のために項目や形式を整えたデータの集まりである。だから、各データを相互に照らし合わせて同一のデータかどうかを特定する（情報の同定）といった操作も比較的容易である。次は、2つあるいはそれ以上のデータベース間での統合である。図らずも情報リンケージ（名寄せ）が有名になった年金納付記録の照合は、各所がもつデータベースの間で個人情報の同定を行う。そして、同一人物と特定できたものを統合する。そのようにしてデータベース全体を統合していく。

ここまでは形式が整ったデータベースでの操作である。しかしデータベース以外にも、ウェブ上などにはたくさんの有用な情報がある。まったく整理されていないこれらの情報をいかに取り入れていくかが、残る課題である。つまり、データベース内とデータベース外との情報リンケージが第3段階、完全にデータベース外での操作が第4段階となる。

具体的なサービスに

この情報リンケージだけでも十分価値があるようにみえるが、プロジェクトの目的は、情報リンケージによって情報を集め、そのうえで新たに生み出される“何か”を有効利用することである。この“何か”を具体的な形にすることを主目的に共同研究を行っているのが、学術コンテンツサービス研究開発センター（センター長・武田英明教授）の蔵川 圭特任准教授である。

「学術をテーマにしたコンテンツサービスの開発を目的に、まず国内の研究者の名寄せを行っています。これは、たとえば共同研究者を探すときや、あるいは研究者の評価などに役立てられるでしょう」と蔵川准教授。文部科学省にある科研費の研究者番号データベースや、各大学の研究者総覧から名寄せを行っており、これは第2段階を具体的に形にしたものといえる。

高久雅生（たかくまさお）・新領域融合研究センター研究員は、同じく科研費のデータベースを中心に、人どうしのつながりに焦点を合わせて再編成する研究を行っている。共同研究者や、研究代表者と分担者の関係などに着目することで、「たとえば、だれをボスにすると研究費が獲得しやすいか、なんてこともわかるかもしれません」と高久研究員。情報リンケージの仕方によって、得られる情報はいろいろ違ってくるようだ。

相澤研究グループでは、海外からのインターン学生も活躍している。Dang Bac Vanさんはベトナム国家大学ホーチミン市校の修士課程の学生で、テキストデータからの情報抽出をテーマに研究している。

これは応用範囲が広い魅力的な研究テーマである。たとえば論文に書かれた手法、ツールなどの名称や関係を抽出することで、そこから他分野への広がりが期待できる。現在は、いかに人手によるコストをかけずにテキストデータを抽出するかというテーマに力を注いでいる。

「情報リンケージ」に集まった仲間

相澤教授を中心とする研究グループの面々は、みなはじめから情報リンケージ研究を行っていたわけではなく、共通点も少ない。相澤教授は通信から始まり、機械学習、最適化、テキスト・自然言語処理など。蔵川准教授は、機械における設計学、ソフトウエアの設計など。相澤教授との共通点は同じデータに触れていることだけ。そんな研究者たちが集まっている。

また、日ごろの研究風景も、同じ部屋で毎日顔を突き合わせている理系の一般的な研究室のイメージとは違い、部屋は全員別々で必要なときに集まり議論する。Vanさんは学生であるため定期的に相澤教授とのディスカッションを行い、「教授からはアイデアをたくさんもらい、充実した研究ができています」というが、基本的には必要なときにのみである。しかし集まったときには、「みな異なるバックグラウンドをもっているので、議論するたびに新たな刺激を受け、とても楽しく研究をしています」（相澤教授）と団結力は堅い。まさに「情報リンケージ（名寄せ）」に“名寄せ”られた研究グループなのだ。

より使いやすく、より価値のあるものを

「現在、第2段階までは問題なく情報リンケージができるシステムを作り出しています。ただ、人手による検証が必要ですので、どの程度の品質でどの程度のコストをかけて行うかという問題になってきます」と相澤教授。研究費が特定の個人に集中する問題や年金など、データベースを共有することで解決する問題は多い。あとは運用方針しだいなのだろう。

今後の方針について、「情報リンケージによって、いままでわからなかったことが断片的に捉えられるところまできています。次は、その断片的な情報の価値を判断して、本当に役に立つ情報を社会に提供できるようにしていく予定です」と相澤教授。一方で教授は、新しいサービスのアイデアも練り始めている。

「情報リンケージ」に集まった研究者たちがどんな新しい価値を生み出していくのか、今後が楽しみである。

（取材・構成　吉戸智明）

Kei Kurakawa
蔵川　圭
学術コンテンツサービス研究開発センター　特任准教授

Masao Takaku
高久雅生
新領域融合研究センター研究員

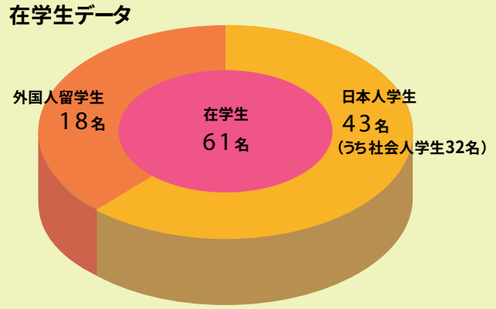
Dang Bac Van
国際インターンシッププログラムによるインターン学生

*1 情報を一カ所に集めるのは同じだが、情報リンケージは、集めた情報間のつながりを重視する。

*2 新領域融合プロジェクト「分野横断型融合研究のための情報空間・情報基盤の構築」には3つのサブプロジェクトがあり、相澤教授のグループはサブプロジェクト1「大規模・異種情報の収集・解析・結合・分類の手法および知識基盤の構築」に含まれる。現在、5年計画の3年目である。

NIIが誇る大学院教育

**情報学の総合的研究拠点であり大学共同利用機関、そして、
情報学分野の高度な人材を育成するという、3つの役割を担う国立情報学研究所（NII）。
その大学院での教育の現状について、学生募集ワーキング主査を務める
曽根原 登教授と卒業生、在学生にざっくばらんに語ってもらった。**



NIIの大学院は、全国に18ある文部科学省の国立研究機関等が共同で設けた教育機関・総合研究大学院大学（総研大）の教育拠点の1つ。情報学専攻の学生約60人が、世界のトップレベルの研究環境の中で学んでいる。

曽根原：皆さんには、NIIの大学院の話をしてもらうために集まってもらいました。まず、ここに進学した理由と自己紹介をしてください。

大向：ブログやソーシャルネットワークが、コミュニケーションをどう変えるかを研究しています。修士までは京都にいましたが、研究に必要な情報量の差を感じて東京に出てきました。

梶山：爆発する情報から、いかに自分に必要な情報を探すかが重要となっています。私は、検索行為そのものが楽しいと感じられるような検索インタフェースの提案を目指していて、情報検索をリードするNIIは、最適な研究場所だと考えました。

カノックワン：私は電子商取引を普及させるうえでの問題点を科学的に解析しています。日本に来て語学を勉強しながら、インターネットで自分の研究テーマに合った進学先を探していて、NIIを見つけました。人の薦めもあって、進学を決めました。

Noboru Sonehara
曽根原 登 教授
情報社会相関研究系 情報制度学部門教授
デジタル商取引における課題解決法や情報信頼基盤の構築などを研究

学び、研究する博士課程

カノックワン：ここは研究設備が新しく、先生もプロフェッショナルぞろいです。気軽に相談できる雰囲気で、研究を進めるうえで不安を感じたことはありません。

梶山：博士課程は学生ですが、多くの大学では後輩の面倒ばかり見なければならず、自分の研究に専念できないと感じます。その点、NIIでは、学生として十分学ばせていただきました。

大向：それに、自由に研究できる雰囲気が好きです。あまりに自由すぎて、ほったらかしと感じる人もいるかも知れませんが、自分でテーマを決め、研究を組み立てたい学生には最高の場所ですね。

曽根原：研究は本来、自由な発想があってこそファイアーするもの。小さい大学院大学だから自由に研究する環境を提供できると自負しています。のびのび研究してもらうために、NIIのネットワークや学術コンテンツを利用して、良い研究成果を、どんどん世の中に出しています。

梶山：NIIには、研究成果を社会に出そうというエネルギーがあり、私は特許を出願していただきました。

大向：私は、在学中にRSSリーダー（Rich Site Summary、ブログを読むためのツール）を普及させようと会社を立ち上げた経験があります。

曽根原：ここは、総合研究大学院大学(総研大)の基盤機関18カ所のなかで、唯一の情報学の総合研究所です。教職員のミッションには、日本の知識・情報サービス戦略や情報通信技術戦略を発信することが含まれている。だから、常に世の中のことを意識しています。

Tomoko Kajiyama
梶山朋子
2007年博士課程修了
現在、早稲田大学人間科学部人間情報科学科助手

information

- **平成20年4月入学 大学院情報学専攻学生募集**
- **募集課程**
  5年一貫制博士課程：学部卒業者及び卒見込者
  博士後期課程：修士課程修了者及び修了見込者
- **出願期間（共通）：平成19年12月14日（金）～12月20日（木）**
- **選抜期日：平成20年2月4日（月）、2月5日（火）のうち1日**
- **入試説明会**
- **開催日：11月6日（火）時間未定**
- **場所：国立情報学研究所**
- **内容：出願に当たっての注意、学生生活紹介等**

＊詳細は、HPでご案内いたします。

NII情報学専攻　検索
http://www.nii.ac.jp/graduate/

多様な教員と学生が生む効果

大向：“人”がバリエーションに富んでいるのも、NIIの特徴です。在学生のうち、留学生が３分の１、社会人学生が２分の１ほどです。これは、ほかの大学院では考えられないバランスです。先生の中にも、企業経験者が多くいますね。

梶山：だから大学ほど閉鎖的ではなく、社会に出る前の私には良い空間でした。カノックワンさんとは席が近かったので、仲良くしていました。お互いに良い刺激を与えていたと思います。

大向：今になって面白いと思うのは、NIIの学生は、全員が“環境を変える”というチャレンジをした人たちだということ。留学生や社会人学生はもちろんですが、博士課程でここを選んだ人も、チャレンジングな人だということです。場所を変えれば、考え方も変わる。テーマを変えることにもつながります。そういう人たちの集まりなので、みんな考え方がシッカリしていますね。

広がれNII遺伝子

大向：ここの教員になって3年ですが、研究のための環境は最高です。ずっといたいと思う一方で、よい研究をするには、常に外の世界とつながっていなくてはなりません。今はNIIを起点に走り回っています。

Ikki Ohmukai
大向一輝
NII第一期生として2005年博士課程修了
現在、国立情報学研究所コンテンツ科学研究系助教

曽根原：卒業生には積極的に社会に出て行って欲しいと思っています。時代は遠心力を求めている。遠心力とは、ここで学んだこと、つまりNIIの遺伝子を世の中に送り出すことです。それが最高学府の使命でもあるわけで・・・。

梶山：私は、この4月から早稲田大学に就職しました。女性ですし、今が自分のためだけに時間が使える貴重な時期だと考えると、研究に集中したいという気持ちもありました。しかし、今は研究しながら、学生に自分の知識を伝える仕事にやりがいを感じています。

大向：ここの卒業生の多くは、学生の時に、共同研究を行った機関などに就職します。それは、学生時代の仕事が認められたからこその結果で、誇れることですね。

カノックワン：私は、卒業したらタイに帰ろうと思っていましたが、今は、NIIでも、日本企業でも、チャンスのある場所で研究を続けたいと考えています。

梶山：ここは研究するには最高の環境ですが、大学院はあまり人に知られていないことが残念ですね。

大向：私は新しいもの好きなので、ネームバリューは自分で作るものだと思っています。自分がやったことが、NIIの歴史をつくっていると実感しています。

曽根原：少子化で、学生の獲得にはどの大学も苦労しています。ここでも、社会人学生や留学生にも学びやすい環境づくりなどに努めています。しかし、NIIのブランド力アップは、今日、集まってくれた皆さんたちが、社会でいかに活躍してくれるかにかかっていると思うのです。研究と大学院教育の両輪がうまく回ってくれたら、NIIは進化します。

忙しいところ今日は、ありがとうございました。

（取材・構成　池田亜希子）

Kanokwan Atchariyachanvanich
アチャリヤチャンワニッ　カノックワン
2001年Asian Institute of Technology修士課程修了 タイからの留学生、後期博士課程3年生

NII ESSAY

「名寄せ」とはどういうことか

相澤彰子（国立情報学研究所 コンテンツ科学研究系教授）

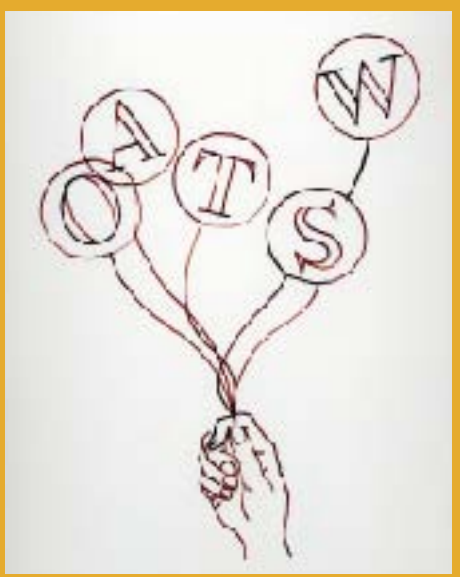
人の一生には色々な出来事がある。例えば、出生・死亡、結婚・離婚といった一般的な事例から養子縁組・移民などがある。しかもこれらは、その都度、届出が必要であり、地理的に分散しているものも多い。また届出には人手が伴うため、内容の綴り違いや入力ミスも起こりうる。

「記録の連結」の難しさ

このように、地理的に分散する個人の記録を、一人一人についてつなぎあわせ、「記録を連結」することは、簡単なようで難しい一方で、社会生活においては最重要である。というのは、本人であることを証明し、家族給付や年金を受け取る根拠を提供する基本データだからだ。

記録の連結とは、二つのデータを付き合わせ、同一内容か否かを判断することである。一見、コンピュータには得意そうに思えるがそうではない。世界最初のフォン・ノイマン型計算機が考案されたのが一九四六年。それ以降、社会のＩＴ化は急激に進んだが、しかし、記録の連結という問題の難しさの本質は変わっていない。

その難しさを示す好例を挙げよう。検索ソフトを使って、ある人物について調べたいとする。まず、「姓名」を入力してみる。同姓同名、地名など、大量のデータが表示される。さらに、所属その他の関連キーワードを追加入力すれば、検索範囲は絞られていく。しかし、そこに記述されている内容が、特定個人の内容か否かの確定には、最終的に人手による検証を待たねばならぬケースが多い。すなわち、記録の連結の役割は、「この記録とこの記録の人物は同じか」、という記述の検証の問題でもある。



記述の検証とコスト

記録の連結の役割が、人物の記述の検証に帰着するとき、近年、年金問題で新聞紙上を騒がせて時代の言葉になった「名寄せ」を思い出さずにはいられない。

「加入したはずの年金が未加入になっている」、「消えた年金は誰のものか」など、人物に関する記述が正しいことを検証し、同一個人として「名を寄せ集める」ことに過ぎない。

しかし、半世紀余を経て解けぬこの難問は、情報処理の本質的難しさを持っているように思われる。すなわち、人手が介在する限り表記の曖昧さは避けられないものであり、曖昧さを前提に情報処理に取り組まねばならない。このとき重要なのは、だれがどのような形で検証したのかという情報、すなわち責任の所在の明確化である。この責任ゆえ、記述の検証には大きなコストがかかるわけだ。

ただし、この場合のコストは本来、情報の不完全さによる損失の穴埋めではなく、「検証」と「責任の所在」の明確化という付加価値への投資とみなすべきであろう。それゆえ、未検証な記述をあたかも検証されたもののように扱ったりすることは、大混乱を招くもとにもなり、年金問題のケースはこれにあたる。

また、自動的な情報処理による劇的なコスト削減という幻想は、半世紀に及ぶ研究が示す通りＩＴ化の価値を過剰に見積もっているに過ぎない。

記録の連結という課題は、計算機で自動獲得された知識と専門家により検証された知識のどちらも財産であり、両者ともに必要であることを示した好例といえよう。

情報から知を紡ぎだす。
NII
国立情報学研究所ニュース（NII Today） 第37号 平成19年9月
発行：大学共同利用機関法人 情報・システム研究機構 国立情報学研究所 http://www.nii.ac.jp/
〒101-8430 東京都千代田区一ツ橋2丁目1番2号 学術総合センター
編集長：東倉洋一 表紙画：小森 誠 写真撮影：由利修一 デザイン：鈴木光太郎 制作：サイテック・コミュニケーションズ
本誌についてのお問合せ：企画推進本部広報普及チーム TEL:03-4212-2135 FAX:03-4212-2150 e-mail: kouhou@nii.ac.jp