慶應義塾大学学術情報リポジトリ
Keio Associated Repository of Academic resouces

Title	SFCにおけるバイオインフォマティクスの歩み： これまでの20年間と今後
Sub Title	Bioinformatics in SFC : the past twenty-years and the future
Author	斎藤, 輪太郎(Saito, Rintaro)
Publisher	慶應義塾大学湘南藤沢学会
Publication year	2015
Jtitle	Keio SFC journal Vol.15, No.1 (2015.) ,p.22- 39
JaLC DOI	10.14991/003.00150001-0022
Abstract	本稿では慶應義塾大学湘南藤沢キャンパス(SFC)におけるバイオインフォマティクスという分野の20年間の開拓史を簡潔に振り返る。特に(1) ゲノム上の境界領域(転写される領域とされない領域の境界等)を認識する分子メカニズムはどのようなものだろうか? (2) コンピュータを用いて如何にゲノムワイドな分子間相互作用ネットワークから新たな生物学的知識を抽出するか? という2点を主なテーマとして採り上げた。この分野が現在抱える問題点や今後の方向性についても触れたい。
Notes	特集 世界を救え：SFCバイオの挑戦#招待論文#第1章 SFCが先導した生命科学
Genre	Journal Article
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=0402-1501-0022

慶應義塾大学学術情報リポジトリ(KOARA)に掲載されているコンテンツの著作権は、それぞれの著作者、学会または出版社/発行者に帰属し、その権利は著作権法によって保護されています。引用にあたっては、著作権法を遵守してご利用ください。

The copyrights of content available on the KeiO Associated Repository of Academic resources (KOARA) belong to the respective authors, academic societies, or publishers/issuers, and these rights are protected by the Japanese Copyright Act. When quoting the content, please follow the Japanese copyright act.

［招待論文］

SFC におけるバイオインフォマティクスの歩み

これまでの 20 年間と今後

Bioinformatics in SFC

The Past Twenty-years and the Future

斎藤 輪太郎
カリフォルニア大学サン・ディエゴ校医学部プロジェクト研究員

Rintaro Saito
Associate Project Scientist, University of California, San Diego, Department of Medicine

Abstract:　本稿では慶應義塾大学湘南藤沢キャンパス (SFC) におけるバイオインフォマティクスという分野の 20 年間の開拓史を簡潔に振り返る。特に (1) ゲノム上の境界領域 (転写される領域とされない領域の境界等) を認識する分子メカニズムはどのようなものだろうか？　(2) コンピュータを用いて如何にゲノムワイドな分子間相互作用ネットワークから新たな生物学的知識を抽出するか？　という 2 点を主なテーマとして採り上げた。この分野が現在抱える問題点や今後の方向性についても触れたい。

Here, 20-years-pioneering-history of bioinformatics at Keio University Shonan Fujisawa Campus (SFC) is briefly reviewed, mainly focusing on two topics: (1) what are the molecular mechanisms that allow recognition of the boundaries of the genomes (boundaries between untranscribed and transcribed regions of the genome, etc)? (2) using computers, how can we efficiently analyze genome-wide molecular interaction network to discover novel biological knowledge? This review may also give readers some ideas of current issues and possible future directions of the field.

Keywords: バイオインフォマティクス、ゲノム、分子間相互作用ネットワーク
bioinformatics, genomes, molecular interaction networks

1　序論

ワトソンとクリックの DNA 二重らせん構造の発見から加速した分子生物学の発展およびそれを研究する実験装置の飛躍的な進歩により、分子生物学的な情報が爆発的に増大した。特に 1990 年代よりヒトの手ではこの膨大な情報

を扱いきれなくなってきた。そしてその頃より、コンピュータで分子生物学的情報を処理する研究すなわちバイオインフォマティクスの必要性が叫ばれた。SFC では 1994 年より分子生物学の分野に踏み込み、バイオインフォマティクスの研究を公式に開始した。当時バイオインフォマティクスの研究を行っている国内の研究機関は少なからずあったものの、そのテーマの多くは分子生物学のための解析手法やツール開発であった。例えば、1995 年に開催されたバイオインフォマティクスのワークショップ“Genome Informatics Workshop 1995 (Dec. 11-12, 1995 Pacific Convention Plaza, Yokohama, Japan)”の 9 件の口頭発表のうち、6 件は DNA やタンパク質の配列類似性解析 (アラインメント) やそれに基づく進化系統樹作成など配列解析手法がテーマであり、残りの 3 件は知識ベースの構築やシミュレーション手法に関するものであった。大雑把に言い換えれば、そこで発表されていたテーマは分子生物学を応用とする「情報科学」であり、得られる結果は主に新規手法の性能ということになる。これに対して当時 SFC のバイオインフォマティクスでは情報科学を用いた「分子生物学」に力を入れた。それが功を奏し、細胞内の遺伝情報の読み取り機構やその進化などの生命現象に関して様々な示唆が得られた。

それから約 20 年、バイオインフォマティクスは大きな発展と変化を遂げたが、その中で常に最先端を目指した SFC のバイオインフォマティクスの軌跡を大雑把に辿ってみたい。

２　ゲノム配列パターンの生物学的意義の解読の試み

遺伝子に関する情報すなわち遺伝情報は DNA という生体内高分子に書き込まれている (コードされている)。特にある生物の DNA にコードされている遺伝子の集合をゲノムという。ゲノムには多くの生命現象を解明する大きな手がかりがあると考えられ、微生物を始めとして、様々な生物のゲノムの読み取りが進められた。しかしゲノム情報自体は A(アデニン)、C (シトシン)、G (グアニン)、T (チミン、RNA では U: ウラシル) の 4 種類の塩基という化学物質からなる膨大な量の暗号 (配列) に過ぎない。従って生物がどのような仕組みで生きているのかを分子レベルから解明するには、ゲノム配列を読むだけでは不十分でその生物学的意味を解読しなければならないのである。ゲ

ノム配列の長さは大腸菌のような比較的単純な生物でも 460 万塩基、ヒトに至っては約 30 億塩基もある。もはや人の手で扱える情報量ではなく、その解読にはコンピュータによる情報処理すなわちバイオインフォマティクスが必要不可欠となる。

様々な生物のゲノム配列が読み取られていく中で、その解読を行う上で多くの研究者が初期に着目したのは、ゲノム配列の中でどこがタンパク質をコードする遺伝子領域なのかということであった (Shepherd JC, 1981; Fickett JW, 1982; Staden R, 1984)。ゲノムにコードされた遺伝情報が細胞内で働くためには、図 1 に示す通り、転写というプロセスを経てゲノム配列の一部が RNA という生体高分子の一種である mRNA に写し取られ、さらに翻訳というプロセスの中で mRNA の一部の配列を元にタンパク質が合成される。そして様々な遺伝子から合成される多種多様なタンパク質が細胞内で酵素として機能したり、栄養を運搬したり、免疫システムを作ったりと生命維持に必要な様々な機能を果たすのである。このようにタンパク質をコードするゲノム上の遺伝子領域をタンパク質コード領域、あるいは短縮してコード領域と呼ぶ。そして図を見れば分かるように、全てのゲノム配列がコード領域というわけではない。例えば、大腸菌ゲノムの数パーセントはコード領域を含まないし、ヒトゲノムに至っては、コード領域はゲノム全体のわずか数パーセントに過ぎないと考えられている。

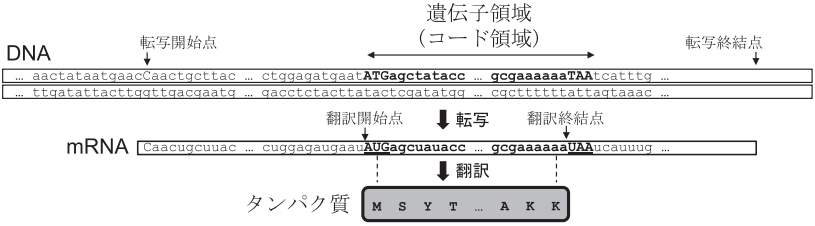

図 1　ゲノム中の遺伝子領域が細胞内で読み取られる仕組み

ゲノムの化学的実体は二本鎖の DNA である。DNA が持つ情報は A,C,G,T の 4 種類の塩基の配列で表される。ゲノムの一部は転写というプロセスを経て RNA という分子に写し取られる。RNA がさらに翻訳というプロセスを経ると、RNA の一部が持つ情報を元にタンパク質が合成される。タンパク質をコードする RNA を mRNA という。タンパク質をコードする DNA や mRNA の遺伝子領域はコード領域と呼ばれる。

ここで我々が着目したのは、細胞内ではゲノム中で転写される部分とそうでない部分、さらに mRNA の中で翻訳される部分とそうでない部分がどのようにして見分けられているのかという点である。そしてその仕組みを解明する鍵は境界領域、すなわちゲノム配列であれば転写開始点と終結点付近、mRNA であれば、翻訳開始点と終結点付近の配列パターンや構造にあるのではないかと考えた。実際、例えば一部の遺伝子の転写開始点の上流にはプロモーターと呼ばれる配列が見られ（次節でさらに触れる）、翻訳開始点であれば、AUG という配列パターン（開始コドン）が翻訳開始シグナルの一部として機能していることが実験的に確かめられている等、特定の生物種の特定の遺伝子群については境界領域に出現するシグナル配列の一部が知られていた。そこで我々はこの知見をさらに多くの生物種で網羅的に確かめるべく、遺伝子データベース GenBank（Burks C. *et al*., 1985）に登録されている様々な種の境界領域を網羅的にかつ様々な角度からコンピュータで解析し、その特徴抽出を行った。その結果、境界領域に顕著に出現する配列パターンや構造は生物種によってかなり異なることが分かってきた。

例えば、転写された mRNA の折れ曲がりの構造（二次構造）である。配列パターンに対して、コンピュータを用いて熱力学的に解析すると、mRNA が細胞内でどのような構造をとるか、どのような折れ曲がり方をするか予測することができる。大腸菌や枯草菌等 4 種の転写終結領域では mRNA が折れ曲がることが配列パターンの熱力学的解析から予測されたが、マイコプラズマ菌など 12 種ではそのような特徴は見られなかった（Washio T. *et al*., 1998）。

同様の解析を rRNA という分子の遺伝子領域の両末端に対しても行った。rRNA はリボソームと呼ばれる細胞内小器官を構成する RNA である。大腸菌の rRNA の遺伝子は DNA から転写された後、（翻訳はされずに）図 2A のように末端同士が対合し、RNase III という酵素によってその構造が認識されて切断（プロセッシング）されることによって、rRNA の最終産物ができることが先行研究で実験によって確かめられていた。我々は他の生物についてもこのようなことが起きているのか、コンピュータ解析によって予測した。その結果、ピロリ菌等、両末端の対合が予測される種もあれば、シアノバクテリア等、対合が明確に予測できない種もあることが分かった（図 2B, Saito R. *et*

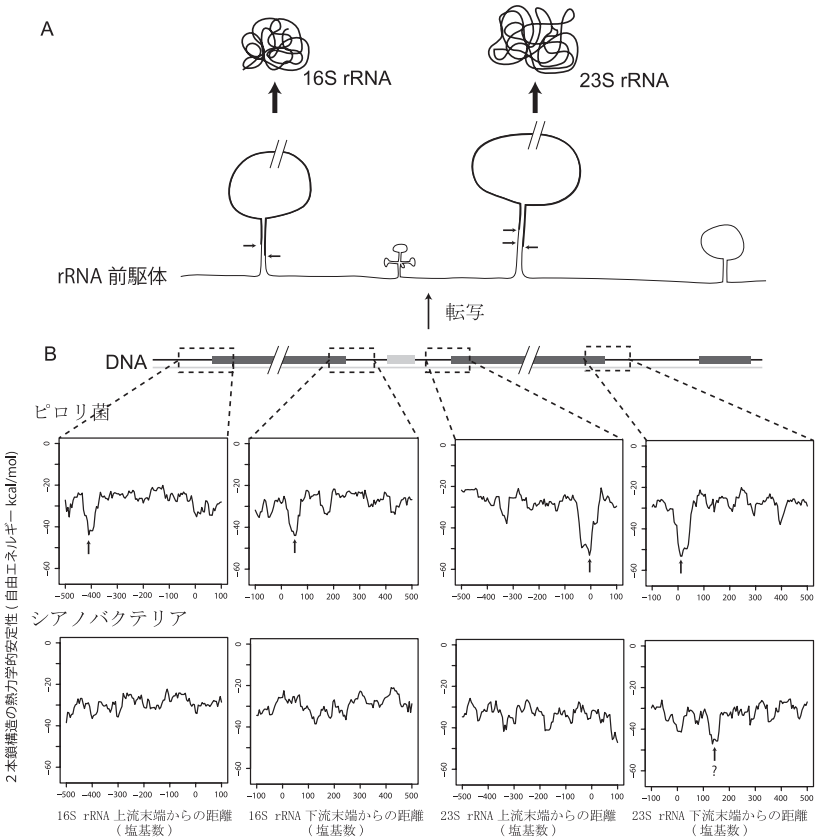

図２　バクテリアの RNA 前駆体の切断

A: DNA から転写された RNA の前駆体はいくつかの RNA 遺伝子を含む。大腸菌では 16S rRNA と 23S rRNA の両端が結合し (ステム構造の形成)、その構造を酵素が認識して切断することによって (切断部位を矢印で表す)、成熟した rRNA が出来上がる。

B: ステム構造の取りやすさの解析。一番左のグラフは、16S rRNA の上流側の末端 (5' 末端) がどれくらい下流側の末端 (3' 末端) と結合しやすいと予測されるかを自由エネルギーという指標で表す。左から 2 番目のグラフは逆に 16S rRNA の下流側の末端がどれくらい上流側の末端と結合しやすいかを表す。同様の解析を 23S rRNA に対しても行った (右半分の 2 つのグラフ)。矢印はステム構造を最も取りやすいと思われる箇所を表す。

al., 2000)。

その他行われた境界領域の解析の主な例を表１に示す。これら一連の解析から分かってきたことは、大腸菌等主要なモデル生物の一部の遺伝子群で確かめられた境界領域認識のための分子メカニズムは他種では必ずしも使われていないようだということである。まさにコンピュータによる網羅的解析が可能になったからこそ得られた推測だと言えよう。では代わりにどのようなメ

表１　ゲノムの一次元配列の解析から得られた境界領域に関する知見

境界領域	着目した生命現象	解析結果、注目すべき点
転写開始部位	エピジェネティクス	CG という配列パターンが多くの種で観測されるが、その頻度は種によってかなり異なる（Shimizu T.S. *et al*., 1997）
転写終結部位	転写終結機構	mRNA の折れ曲がり（ステムループ形成）による転写終結を行っていると思われる種とそうでない種がいる（Washio T. *et al*., 1998）
翻訳開始部位	翻訳開始機構	古細菌の翻訳開始機構は真正細菌と真核生物の両方の特徴を併せ持つことが予測される（Saito R. *et al*., 1999）
		mRNA の上流が 16S rRNA と強く結合すると予測される種とそうでない種がいる（Osada Y. *et al*., 1999）
	翻訳量制御	タンパク質合成量が高いと思われる遺伝子の翻訳開始領域にはより顕著な配列パターンが観測される傾向がある（Sakai H. *et al*., 2001）
	RNA 分解制御	翻訳開始領域上流に短いタンパク質コード領域があると、RNA の分解が促進されると予測される（Matsui M. *et al*., 2007）
翻訳終結部位	翻訳終結	翻訳終結部位付近に特徴のある配列パターンが観測されるが、これは生物によって異なる（Ozawa Y. *et al*., 2002）
	終止コドンのリードスルー	ショウジョウバエの遺伝子の一部で本来の翻訳終結部位の読み飛ばしが起こっている（Sato M. *et al*., 2003）
RNA 切断部位	RNA のプロセッシング	rRNA の両末端が結合すると予測される種と、そうでない種がいる（Saito R. *et al*., 2000）
	スプライシング	恒常的に切断される場合と、状態によって切断される場合で RNA 上のシグナル配列が異なる（Itoh H. *et al*., 2004）
反復配列の開始部位	ゲノム進化	TTTTAAAAA という配列パターンがゲノム中の反復配列の一種である Alu の上流に観測される（Toda Y. *et al*., 2000）

カニズムが使われているのだろうか。当時はっきりしたことはまだ分からなかったものの、我々は境界領域の認識に関わっている可能性のある新たな配列パターンをいくつか発見している。例えば、情報理論を応用した配列解析を行うことにより、多くの種の翻訳開始部位直前の 2 つの塩基の種類には強い相関が見られることが判明し (Osada Y. *et al*., 2006)、哺乳類の翻訳終結部位直後の 1 塩基は A や G になることが多いことが分かった (Ozawa Y. *et al*., 2002)。

さらに実験技術が進んで細胞内に関する多様なデータが入手可能になったことにより、細胞内で境界領域が識別される仕組みを考える上で単なるゲノム一次元配列だけでは情報として必ずしも十分ではないことがより鮮明になってきた。これについては次節で解説しよう。

3　DNA に結合しているヒストンの情報の統合

前節で述べた解析手法では主な解析対象は A,C,G,T の 4 種類の塩基からなるゲノムの 1 次元配列情報である。以前は特にバクテリアでは転写開始領域付近に存在するプロモーターと呼ばれる配列が転写開始を決める重要な因子の 1 つとして着目されていた (Mulligan M.E. *et al*., 1984)。しかし実際には

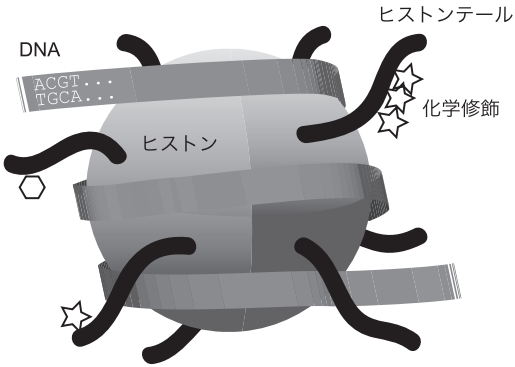

図 3　ヒストンとそれに巻き付く DNA 構造の簡略化した模式図

図で示したヒストン (コアヒストン) は H2A, H2B, H3, H4 という 4 種類のヒストンタンパク質が 2 個ずつ集まって複合体を形成しており、そこに 2 本鎖の DNA が巻きついている。ヒストンの末端であるヒストンテールは化学修飾を受け、これがしばしば転写制御に使われる。

図３に示すように、ゲノムの化学的実体である DNA は複雑な３次元構造をとっている。真核生物（植物、動物等細胞核を持つ生物。バクテリアは真核生物ではない）では DNA はそれ単独で細胞内に存在しているわけではなく、ヒストンと呼ばれるタンパク質に巻き付いており、ヌクレオソームと呼ばれる構造を形成している。さらに複数のヌクレオソームやそれに結合するタンパク質が集まってクロマチンと呼ばれる高次構造が観測されるが、このクロマチンが凝縮していると、一般的にそこにあるゲノム領域からの転写が抑制される。またヒストンの一部であるヒストンテールと呼ばれる部分が化学修飾を受ける事により、転写が活性化されたり、また逆に抑制されたりする。すなわちあるゲノム領域から転写が起こるか否か、あるいは生体内でどのように遺伝子領域が見分けられているのかを考えるとき、ゲノムの一次元配列だけでは情報として必ずしも十分ではなく、DNA と他の分子が形成する高次構造や化学修飾をも考慮する必要がある。どのようなヒストン修飾があると転写が活性化されたり、逆に抑制されるのか研究が進められており、例えば、転写量が多い遺伝子の転写開始部位付近の特定のヒストンテールの特定の部分 (H3K4 と呼ばれる部分) は多くの場合、3 重にメチル化（メチル基 -CH_3 が結合する化学修飾）されている。つまり、生体内ではヒストン修飾パターンとゲノム配列パターンの両方が認識されて転写が行われていることが示唆される。

それではヒストン修飾パターンとゲノム配列パターンは生体内では互いに独立して機能し、あるいは独立に進化してきたのか。それとも両者は何らかの形で互いに関連しているのか。我々はこのヒントを探るべく、ゲノム配列パターンとヒストン修飾パターンの間に何らかの相関はないか、コンピュータ解析を行った（Nozaki T. *et al*., 2011）。解析対象としたのは、CG という配列パターンを持ち、複数の箇所からの転写を促進する“ブロード型プロモーター”を持つ遺伝子領域と、TATA という配列パターンを持ち、一定の箇所からの転写を促進する“ピーク型プロモーター”を持つ遺伝子領域である（図 4A）。その結果、ブロード型プロモーターを持つ遺伝子の方がピーク型プロモーターを持つ遺伝子に比べ、転写開始領域付近の決まった位置に修飾されたヒストンが多く存在していることが判明した（図 4B）。さらにブロード型プロ

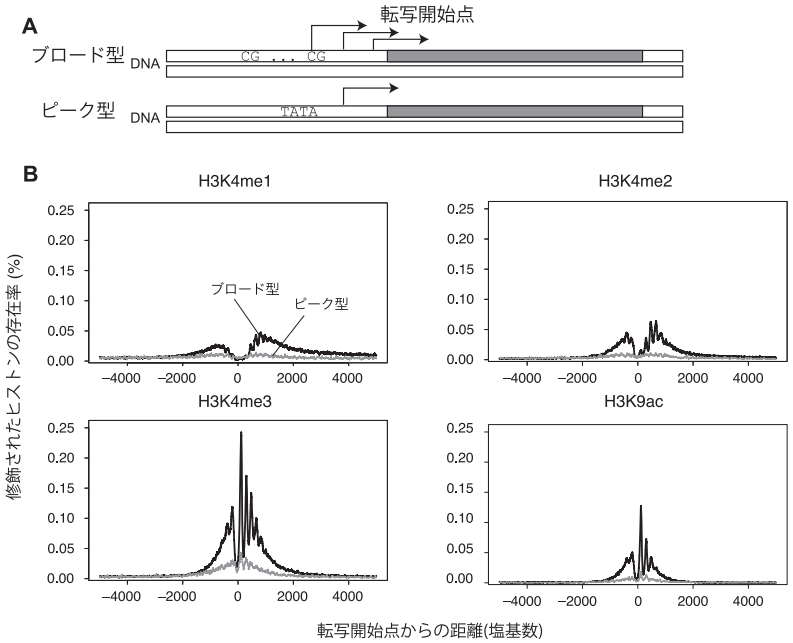

図４　プロモータのタイプと修飾されたヒストンの存在率

A: 我々が解析の対象としたのは、ブロード型プロモーターまたはピーク型プロモーターを上流に持つ遺伝子群である。ブロード型のプロモーターは CG を多く含み、転写開始点のゆらぎが大きい。ピーク型プロモーターは TATA という配列パターンを含み、転写開始点のゆらぎが小さい。

B: 転写開始点付近の化学修飾されたヒストンの存在率。横軸は転写開始点からの距離、縦軸は対応する位置にどれくらいの割合で修飾されたヒストンが存在するかを表す。本解析では 4 種類のヒストン修飾（H3K4me1, H3K4me2, H3K4me3, H3K9ac）を対象とし、存在率の計算はブロード型プロモーターを持つ遺伝子群とピーク型プロモータを持つ遺伝子群に分けて行った。

モータを持つ遺伝子の発現量は付近のヒストンが修飾されているときに転写量が高くなる傾向があることが分かり、ヒストン修飾とゲノム配列パターンの関連が示唆された。もし、ヒストン修飾の認識機構とゲノム配列パターンの認識機構が密接にリンクしているとすれば、ゲノム配列パターンが分子生物学的に意味するところもヒストン修飾の状態によって大きく変化するかも知れず、もはやゲノム配列だけから転写制御を考える時代は終わりつつあるのかも知れない。そして近年では、エンハンサーと呼ばれる遺伝子から遠く

離れたゲノム上の因子が転写に影響することも詳しく調べられるようになった (He B. *et al*., 2014; Zabidi M.A. *et al*., 2015)。従って境界領域の認識等の生命現象を調べる上で、今後はゲノム配列と他の様々な因子がどのように関連してどのような生命現象を引き起こすのか、実験技術と情報解析を駆使したさらに多次元的な解析が必要となっていくだろう。

４　分子間の相互作用の解析

細胞の中には遺伝子がコードされている DNA や RNA 等の核酸、タンパク質、脂質、糖質など様々な分子が存在している。それらの分子は単独で細胞の中で機能するのではなく、他の分子と相互作用する事によって機能を発揮している。ゲノム配列を全部読み取る目標の１つは、全遺伝子や全タンパク質の情報を取り揃えることである。しかし、生命現象を理解するためには、個々の“部品”の情報を揃えるだけでは不十分で、どの部品とどの部品が組み合わさって機能を果たしているのかを探る事が不可欠である。2000 年には様々なタンパク質がどのタンパク質と結合するのかを高速（ハイスループット）に調べる技術（酵母ツーハイブリッド法）が発達し (Ito T. *et al*., 2001)、大量のタンパク質間相互作用情報が得られるようになってきた。分子同士のつながりの集合をネットワークと呼ぶが、SFC では 2002 年よりこれら大量の分子間相互作用ネットワークを解析する研究に着手した。

ハイスループットな実験系から得られる分子間相互作用ネットワークは、一般にとても複雑である(図 5A)。これは解析対象の分子数が増えるに従って、その間の相互作用の数が急増するのが原因の一つである。さらに実験で得られた相互作用が必ずしも実際に細胞内で起こっている相互作用を反映しているわけではない。まず厄介なのが、実験データには多くのノイズが含まれることである (斎藤ら , 2004)。さらに例えば A と B という 2 つのタンパク質が結合することが実験で確認できたとしても、実際の細胞内では隔離されていて結合しない可能性がある。また実験結果から、A-B-C という 3 つのタンパク質の相互作用が得られたとしても、肝臓組織の細胞では A と B、腎臓組織の細胞では B と C が結合し、A-B-C という同時結合は実際の生体内では起こらないということもありうる。そこで、得られた分子間相互作用データを

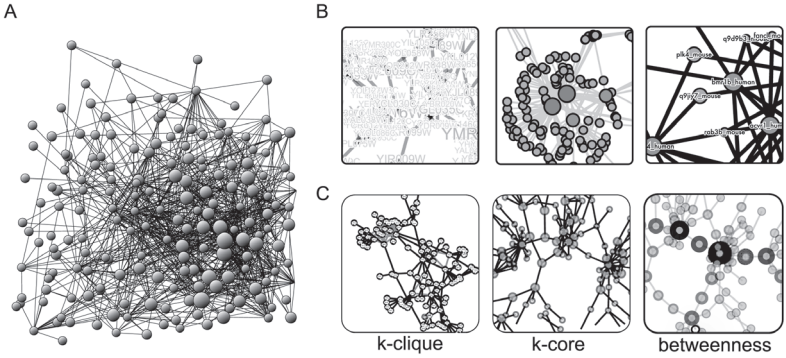

図５　分子間相互作用ネットワークの可視化と解析

A: 細胞内分子間相互作用ネットワークの一部の例。円（ノード）はタンパク質を表し、線（エッジ）は結ばれたタンパク質の間に相互作用があることを表す。

B, C: eXpanda を使った分子間相互作用ネットワークの解析・可視化の例。(B) 様々な可視化。左から順に、ノードを遺伝子名で表示、多数の他の分子と相互作用する（重要度の高いと思われる）分子の強調、ヒトとマウスの相互作用ネットワークの統合可視化。(C) 相互作用ネットワークからの特徴的な部分の抽出。k-clique と k-core は密な部分を抽出する指標。betweeness は与えられたノードがネットワークの中でどれくらい”中心的”であるかを表す指標。詳しくは Wang J. *et al*., 2010 等を参照。

分子生物学的に解釈し、新たな知見を得るためにはまず、注目している生命現象に合致すると推定される相互作用を抽出しなければならない。その主な方法の１つは、実際の生体内の相互作用に有意に見られる特徴をまず列挙し、次に実験から得られた相互作用データの中からその特徴を持つものを抽出することである。例えば、細胞内で複数のタンパク質が安定的に結合して機能している場合（タンパク質の複合体）、そのタンパク質同士の間には密な結合が存在しやすいことが知られている。これをもとに、相互作用ネットワークの中から、グラフ理論に基づいて密な部分を抽出し、新規タンパク質複合体を予測する方法が開発された（Bader G.D. and Hogue C.W., 2003; van Dongen S. and Abreu-Goodger C., 2012）。我々はこれら既存の方法に改良を加え、密な結合に含まれるタンパク質の一部（ドメイン）がどのような組み合わせで結合しうるか、整数計画法と呼ばれる手法を用いて予測する手法を開発した (Ozawa Y. *et al*., 2010)。これにより、どのタンパク質群が複合体を形成するか、ということだけではなく、それらタンパク質のどの部分とどの部分が結合す

るのかを予測できるようになった。

このように相互作用データを与えれば自動的に生物学的興味の対象となり得る相互作用を抽出するシステムの開発は大量の相互作用データを処理する上で必要だが、抽出した相互作用の生物学的意味づけを最終的に行うのはヒトである。例えば、新規タンパク質複合体が予測できれば、その生物学的機能はどんなものか、複合体を構成する個々のタンパク質の機能やその複合体自体が相互作用する他のタンパク質の情報から、それらのタンパク質の機能に関する過去の研究を参照しつつ、考察することになる。つまりどうしても手動で行わなければならない部分が出てくる。そこで、相互作用データを様々な角度から解析し、その結果を可視化・閲覧可能なシステム eXpanda を開発した（図 5B, Negishi Y. *et al*., 2007）。これによってユーザは任意の相互作用実験データを読み込み、生物学的に関係がありそうな相互作用を情報科学的手法で絞込み、ネットワーク描画方法を選択して分かりやすく可視化することができるようになった。

相互作用の中で分子生物学的に特に面白いのはシグナル伝達経路を構成するものであろう。細胞は外界の環境に適応するために、細胞膜のタンパク質（レセプター）から外界の情報を受け取り、タンパク質同士の相互作用や移行、タンパク質の化学修飾（リン酸化）等によってあたかも伝言ゲームのように DNA まで情報を伝え、必要な遺伝子の転写を行っている。この“伝言”がどこかで狂えば、細胞の異常な挙動につながる。例えば、細胞分裂のシグナルが異常をきたせば、細胞分裂が止まらなくなり、腫瘍形成が起こる。そこで実験から得られた大量の相互作用データから新規シグナル伝達経路を予測する研究が少なくとも十数年前から始まっているが（Steffen M. *et al*., 2002）、実験データに含まれるノイズの量の多さ、実験で検出が困難な相互作用の存在、そしてそれをカバーできるだけの情報科学的手法が未開発であることなどが原因で未だ試験的で精度が限られている。我々はタンパク質の化学修飾（リン酸化）が時間とともにどのように変化するかを観測したデータを用いれば、シグナル伝達経路を予測できる可能性を示したが（Imamura H. *et al*., 2010）、今後は細胞内で起きている様々な現象を高い精度で観測可能な実験技術とそれを効果的に処理する情報科学的手法の両方が向上し、高い精度で予測ができる

ようになることが期待される。

5　むすび

本稿は 20 年間の SFC のバイオインフォマティクスの軌跡をまとめるという大きな目標を掲げて書き始められたが、SFC のバイオインフォマティクスの研究成果は論文の数にして既に 100 を超えるため、筆者の力が及ばず、ごく一部をカバーするにとどまり、さらに本稿の内容の半分くらいは筆者が SFC 在籍中に（～ 2011 年）何らかの形で関わったものに偏ってしまった。本稿で触れた以外にも注目すべきバイオインフォマティクスの研究成果が多数あることをご了承頂きたい。また分子生物学を専門としない読者にも興味を持って頂くように構成したため、必ずしも学術的に正確でない記述や飛躍した論理展開があることも否めない。この点についてはご批判を頂くことがあれば、真摯に受け止めたいと考えている。

SFC のバイオインフォマティクスの業績は他の研究機関との共同研究によるものも少なくない。二つ例を挙げると、理化学研究所では 2000 年から網羅的に哺乳類（マウス）遺伝子の機能解析をするプロジェクト FANTOM (Functional Annotation of Mammalian (Mouse) Genome) が始まっているが、SFC は FANTOM 1(Kawai J. *et al*., 2001) と FANTOM2 (Okazaki Y. *et al*. 2002) に参加している。そして FANTOM 1 で得られたマウスの mRNA 配列データを使い（正確には cDNA）、イントロンと呼ばれる mRNA から除去される部分が遺伝子の上流側に偏っていることを示してその機能的・進化的な意味の考察を行い注目を集めた (Sakurai A. *et al*., 2002; Mourier T. and Jeffares D.C., 2003)。2006 年には本塾理工学部と文部科学省のゲノムネットワークプロジェクトに参加し、転写を制御するタンパク質（転写因子）複合体の大規模解析をバイオインフォマティクスの面からサポートした (Miyamoto-Sato E. *et al*., 2010)。

これまでの研究を振り返る中で、あともう一歩だった研究についても少しだけ例を挙げて触れておこう。我々はショウジョウバエで翻訳終結部位で翻訳が止まらない現象（終止コドンのリードスルー）が頻繁に起きているという仮説をコンピュータ解析から立てた (Sato M. *et al*., 2003)。残念ながら、当時

の入手可能なデータの量および質の問題や我々が考案したコンピュータによる解析手順が十分に洗練されていなかったこと、実験による検証までできなかったことなどにより、十分に評価されるに至らなかった。しかし我々の仮説はその 8 年後、他の研究機関により裏付けられた (Jungreis I. *et al*., 2011)。また、本稿で触れた eXpanda (Negishi Y. *et al*., 2007) については、普及を目指して学会発表を精力的に行ったが、カリフォルニア大学サン・ディエゴ校が開発した競合ソフト Cytoscape に遥かに及んでいないことは認めざるを得ない。但し、eXpanda を実際に開発した学生達の名誉のために述べておくと、筆者はその後 Cytoscape を開発している研究室に 3 年間在籍する機会があったのだが、そこで詳しくわかったのは、同研究室では米国政府からの多額の予算をもとに開発スタッフを 4 人以上揃え、他にカリフォルニア大学サン・フランシスコ校やトロント大学等様々な研究機関を連携開発拠点として Cytoscape の多数の機能の実装を行っているということである (斎藤・大野 , 2013)。このような開発の規模の圧倒的な差を考えれば、学生達数名で健闘したと言えるだろう。

本稿ではどちらかというと基礎研究的な側面を強調したが、医学や医療に関係した研究も進められている。例えば我々は理化学研究所や筑波大学等と大腸癌組織の共同研究を行い、正常組織と比較して、癌組織で多く転写される RNA を 71 個発見した (Saito R. *et al*., 2011)。これらは遺伝子の“裏”(遺伝子をコードする DNA 鎖の逆鎖) から転写されるという興味深い RNA (アンチセンス RNA) でこれらの一部が細胞の癌化に関わっている可能性が示唆された。

今後医療や農業分野等からの需要や新たな実験技術の出現、新規生命現象の発見によって、バイオインフォマティクスが対象とする主な分野は大きく変わっていくことが予想される (これまでもそうであったように)。しかしライフサイエンスの研究が続く限り、バイオインフォマティクスは今後も必要不可欠な分野であり続けるだろう。その中で SFC のバイオインフォマティクスがどの分野で活躍し大きく貢献していくのかを期待とともに楽しみにしている。

謝辞

本編でも触れたが、バイオインフォマティクスの分野は移り変わりが激しい。次々と困難な壁がやってくるが、さてこれをどうやって乗り越えようかというとき、本稿を執筆することになり、そもそもこれまでどのように壁を乗り越えてきたのか、振り返るちょうど良い機会となった。今思えば、バイオインフォマティクスを始めた当初は SFC には実験設備も分子生物学の授業も全くないという困難な状況であったが、そのような中で研究室の同僚たちと切磋琢磨しながら何度も議論を重ね、分子生物学に少しでも貢献できるような成果を出し、学会等で発表をしてきた。2000 年から 2 年間理化学研究所に勤務したが、そこで多くの研究者から貴重なことを学んだ。その経験を生かし、2002 年から同僚や学生たちと再び SFC にて研究を始めたが、そこでも多くの同僚に助けられ、また多くの学生の若いパワーで多数の成果を出すことができた。そしてこれは共同研究をさせて頂いた多数の研究機関の方々のご支援があったことも非常に大きい。これらの多くの方々に深く感謝したい。

ところで SFC が分子生物学を始めた 1994 年にはコンピュータで進化をシミュレーションするという、今考えればとても大胆 (無謀 ?) なことをやっていた。だが思い返せば、それも若い学生のうちしかできない貴重な経験であり、当時ともに試行錯誤して頑張った学生時代の研究室の仲間たちにも感謝したい。

そして最後に私の今の米国生活を忍耐強く支えてくれている家族にも感謝する。

参考文献

(SFC の業績、または SFC が関わった業績)

Imamura H., Yachie N., Saito R., Ishihama Y., Tomita M., “Towards the systematic discovery of signal transduction networks using phosphorylation dynamics data.” *BMC Bioinformatics*, 11(1), 2010, p.232.

Itoh H., Washio T., Tomita M., “Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukary-otes.” *RNA*, 10(7), 2004, pp.1005-1018.

Kawai J., Shinagawa A., Shibata K., Yoshino M., Itoh M., Ishii Y., Arakawa T., Hara A., Fukunishi Y., Konno H., Adachi J., Fukuda S., Aizawa K., Izawa M., Nishi K., Kiyosawa H., Kondo S., Yamanaka I., Saito T., Okazaki Y., Gojobori T., Bono H., Kasukawa T., Saito R., Kadota K., Matsuda H., Ashburner M., Batalov S., Casavant T., Fleischmann W., Gaasterland T., Gissi C., King B., Kochiwa H., Kuehl P., Lewis S., Matsuo Y., Nikaido I., Pesole G., Quackenbush J., Schriml L.M., Staubli F., Suzuki R., Tomita M., Wagner L., Washio T., Sakai K., Okido T., Furuno M., Aono H., Baldarelli R., Barsh G., Blake J., Boffelli D., Bojunga N., Carninci P., de Bonaldo M.F., Brownstein M.J., Bult C., Fletcher C., Fujita M., Gariboldi M., Gustincich S., Hill D., Hofmann M., Hume D.A., Kamiya M., Lee N.H., Lyons P., Marchionni L., Mashima J., Mazzarelli J., Mombaerts P., Nordone P., Ring B., Ringwald M., Rodriguez I., Sakamoto N., Sasaki H., Sato K., Schonbach C., Seya T., Shibata Y., Storch K.F., Suzuki H., Toyo-oka K., Wang K.H., Weitz C., Whittaker C., Wilming L., Wynshaw-Boris A., Yoshida K., Hasegawa Y., Kawaji H., Kohtsuki S., Hayashizaki Y., “Functional annotation of a full-length mouse cDNA collection.” *Nature*, 409(6821), 2001, pp.685-690.

Matsui M., Yachie N., Okada Y., Saito R., Tomita M., “Bioinformatic analysis of post-

transcriptional regulation by uORF in human and mouse." *FEBS Lett.*, 581(22), 2007, pp.4184-4188.

Miyamoto-Sato E., Fujimori S., Ishizaka M., Hirai N., Masuoka K., Saito R., Ozawa Y., Hino K., Washio T., Tomita M., Yamashita T., Oshikubo T., Akasaka H., Sugiyama J., Matsumoto Y., Yanagawa H., "A comprehensive resource of interacting protein regions for refining human transcription factor networks." *PLoS One*, 5(2), 2010, p.e9289.

Negishi Y., Nakamura H., Yachie N., Saito R., Tomita M., "eXpanda: an integrated platform for network analysis and visualization." *In Silico. Biol.*, 7(2), 2007, pp.141-144.

Nozaki T., Yachie N., Ogawa R., Kratz A., Saito R., Tomita M., "Tight Associations between Transcription Promoter Type and Epigenetic Variation in Histone Positioning and Modification." *BMC Genomics*, 12(1), 2011, p.416.

Okazaki Y., Furuno M., Kasukawa T., Adachi J., Bono H., Kondo S., Nikaido I., Osato N., Saito R., Suzuki H., Yamanaka I., Kiyosawa H., Yagi K., Tomaru Y., Hasegawa Y., Nogami A., Schonbach C., Gojobori T., Baldarelli R., Hill D.P., Bult C., Hume D.A., Quackenbush J., Schriml L.M., Kanapin A., Matsuda H., Batalov S., Beisel K.W., Blake J.A., Bradt D., Brusic V., Chothia C., Corbani L.E., Cousins S., Dalla E., Dragani T.A., Fletcher C.F., Forrest A., Frazer K.S., Gaasterland T., Gariboldi M., Gissi C., Godzik A., Gough J., Grimmond S., Gustincich S., Hirokawa N., Jackson I.J., Jarvis E.D., Kanai A., Kawaji H., Kawasawa Y., Kedzierski R.M., King B.L., Konagaya A., Kurochkin I.V., Lee Y., Lenhard B., Lyons P.A., Maglott D.R., Maltais L., Marchionni L., McKenzie L., Miki H., Nagashima T., Numata K., Okido T., Pavan W.J., Pertea G., Pesole G., Petrovsky N., Pillai R., Pontius J.U., Qi D., Ramachandran S., Ravasi T., Reed J.C., Reed D.J., Reid J., Ring B.Z., Ringwald M., Sandelin A., Schneider C., Semple C.A., Setou M., Shimada K., Sultana R., Takenaka Y., Taylor M.S., Teasdale R.D., Tomita M., Verardo R., Wagner L., Wahlestedt C., Wang Y., Watanabe Y., Wells C., Wilming L.G., Wynshaw-Boris A., Yanagisawa M., Yang I., Yang L., Yuan Z., Zavolan M., Zhu Y., Zimmer A., Carninci P., Hayatsu N., Hirozane-Kishikawa T., Konno H., Nakamura M., Sakazume N., Sato K., Shiraki T., Waki K., Kawai J., Aizawa K., Arakawa T., Fukuda S., Hara A., Hashizume W., Imotani K., Ishii Y., Itoh M., Kagawa I., Miyazaki A., Sakai K., Sasaki D., Shibata K., Shinagawa A., Yasunishi A., Yoshino M., Waterston R., Lander E.S., Rogers J., Birney E., Hayashizaki Y., "Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs." *Nature*, 420(6915), 2002, pp.563-573.

Osada Y., Saito R., Tomita M., "Comparative analysis of base correlations in 5' untranslated regions of various species." *Gene*, 375, 2006, pp.80-86.

Osada Y., Saito R., Tomita M., "Analysis of base-pairing potentials between 16S rRNA and 5' UTR for translation initiation in various prokaryotes." *Bioinformatics*, 15(7-8), 1999, pp.578-581.

Ozawa Y., Hanaoka S., Saito R., Washio T., Nakano S., Shinagawa A., Itoh M., Shibata K., Carninci P., Konno H., Kawai J., Hayashizaki Y., Tomita M., "Comprehensive sequence analysis of translation termination sites in various eukaryotes." *Gene*, 300(1-2), 2002, pp.79-87.

Ozawa Y., Saito R., Fujimori S., Kashima H., Ishizaka M., Yanagawa H., Miyamoto-Sato E., Tomita M., "Protein complex prediction via verifying and reconstructing

the topology of domain-domain interactions." *BMC Bioinformatics*, 11, 2010, p.350.
Saito R., Kohno K., Okada Y., Osada Y., Numata K., Kohama C., Watanabe K., Nakaoka H., Yamamoto N., Kanai A., Yasue H., Murata S., Abe K., Tomita M., Ohkohchi N., Kiyosawa H., "Comprehensive Expressional Analyses of Antisense Transcripts in Colon Cancer Tissues Using Artificial Antisense Probes." *BMC Medical Genomics*, 4(1), 2011, p.42.
Saito R., Ozawa Y., Kuzuno N., Tomita M., "Computer analysis of potential stem structures of rRNA operons in various procaryote genomes." *Gene*, 259(1-2), 2000, pp.217-222.
Saito R., Tomita M., "Computer analyses of complete genomes suggest that some archaebacteria employ both eukaryotic and eubacterial mechanisms in translation initiation." *Gene*, 238(1), 1999, pp.79-83.
Sakai H., Imamura C., Osada Y., Saito R., Washio T., Tomita M., "Correlation between Shine--Dalgarno sequence conservation and codon usage of bacterial genes." *J. Mol. Evol*., 52(2), 2001, pp.164-170.
Sakurai A., Fujimori S., Kochiwa H., Kitamura-Abe S., Washio T., Saito R., Carninci P., Hayashizaki Y., Tomita M., "On biased distribution of introns in various eukaryotes." *Gene*, 300(1-2), 2002, pp.89-95.
Sato M., Umeki H., Saito R., Kanai A., Tomita M., "Computational analysis of stop codon readthrough in D. melanogaster." *Bioinformatics*, 19(11), 2003, pp.1371-1380.
Shimizu T.S., Takahashi K., Tomita M., "CpG distribution patterns in methylated and non-methylated species." *Gene*, 205(1-2), 1997, pp.103-107.
Toda Y., Saito R., Tomita M., "Characteristic sequence pattern in the 5- to 20-bp upstream region of primate Alu elements." *J. Mol. Evol*., 50(3), 2000, pp.232-237.
Washio T., Sasayama J., Tomita M., "Analysis of complete genomes suggests that many prokaryotes do not rely on hairpin formation in transcription termination." *Nucleic Acids Res*., 26(23), 1998, pp.5456-5463.

(他の研究機関からの論文)
斎藤 輪太郎・鈴木 治和・冨田 勝「ゲノムワイドデータの精製」『蛋白質・核酸・酵素』49(17)、2004 年、pp.2882-2888。
斎藤 輪太郎・大野 圭一朗「Cytoscape による細胞内分子間相互作用ネットワークの解析」『実験医学』31(14)、2013 年、pp.2291-2297。

Bader G.D., Hogue C.W. "An automated method for finding molecular complexes in large protein interaction networks." *BMC Bioinformatics*, 4:2, 2003.
Burks C., Fickett J.W., Goad W.B., Kanehisa M., Lewitter F.I., Rindone W.P., Swindell C.D., Tung C.S., Bilofsky H.S., "The GenBank nucleic acid sequence database." *Comput. Appl. Biosci*., 1(4), 1985, pp.225-233.
van Dongen S., Abreu-Goodger C., "Using MCL to extract clusters from networks." *Methods Mol. Biol*., 804, 2012, pp.281-295.
Fickett J.W., "Recognition of protein coding regions in DNA sequences." *Nucleic Acids Res*., 10(17), 1982, pp.5303-5318.
He B., Chen C., Teng L., Tan K., "Global view of enhancer-promoter interactome in human cells." *Proc. Natl. Acad. Sci. U S A*., 111(21), 2014, pp.E2191-2199.
Ito T., Chiba T., Ozawa R., Yoshida M., Hattori M., Sakaki Y., "A comprehensive two-hybrid analysis to explore the yeast protein interactome." *Proc. Natl. Acad. Sci. U S*

A., 98(8), 2001, pp.4569-4574.

Jungreis I., Lin M.F., Spokony R., Chan C.S., Negre N., Victorsen A., White K.P., Kellis M., “Evidence of abundant stop codon readthrough in Drosophila and other metazoa.” *Genome Res*., 21(12), 2011, pp.2096-2113.

Mourier T., Jeffares D.C., “Eukaryotic intron loss.” *Science*, 300(5624), 2003, p.1393.

Mulligan M.E., Hawley D.K., Entriken R., McClure W.R., “Escherichia coli promoter sequences predict in vitro RNA polymerase selectivity.” *Nucleic Acids Res*., 12(1 Pt 2), 1984, pp.789-800.

Shepherd J.C., “Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification.” *Proc. Natl. Acad. Sci. U S A*., 78(3), 1981, pp.1596-1600.

Staden R., “Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes.” *Nucleic Acids Res*., 12(1 Pt 2), 1984, pp.551-567.

Steffen M., Petti A., Aach J., D’haeseleer P., Church G., “Automated modelling of signal transduction networks.” *BMC Bioinformatics*, 3:34, 2002.

Wang J., Li M., Deng Y., Pan Y., “ Recent advances in clustering methods for protein interaction networks.” *BMC Genomics*, 11 Suppl 3, 2010, p.S10.

Zabidi M.A., Arnold C.D., Schernhuber K., Pagani M., Rath M., Frank O., Stark A., “Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation.” *Nature*, 518(7540), 2015, pp.556-559.

〔受付日　2015. 2. 26〕