Review Article

完全長 cDNA とゲノム医科学

菅野純夫

東京大学医科学研究所 ヒトゲノム解析センターゲノム構造解析分野

Full-length cDNA and Genome Medicine

As the result of the human genome project, the entire sequence of human genome was determined. Unfortunately, it is still not a trivial task to identify genes from these genome sequences only by the computational methods. Thus, full-length cDNAs are indispensable for the identification of genes and the determination of their structures. In addition, cDNAs will be a valuable resource for the functional analysis of the gene. We have performed so called FLJ project of collecting and determining the entire sequences of putative full-length cDNA clones. In this project, we extensively used cDNA libraries made by oligo-capping method. In this review, I will describe the current status of FLJ project and high-through-put functional analysis using full-length cDNA clones obtained by FLJ project.

Rec.4/11/2003, pp157-163

Sumio Sugano

Human Genome Center,

The Institute of Medical Science,

The University of Tokyo

Key words genome, cDNA, high-through-put, functional analysis, full-length

遺伝子と疾患

疾患は環境要因と遺伝要因が複雑に絡み合って発症する. 外傷や中毒のように環境要因がほぼ100%の場合もあれば, 遺伝病のように遺伝要因がほぼ100%である場合もある. しかし, 多くの疾患はこの中間にあり, 現代社会で問題とされる「ありふれた疾患」, がん, 高血圧, 糖尿病, 喘息・アレルギーなども, 体質として漠然と把握されている遺伝要因に栄養, ストレス, 感染, 運動等の環境要因が働いて発症・進展すると考えられている.

遺伝病での遺伝要因は, 多くの場合一つの遺伝子の変異であり, 家系を用いて比較的容易に解析できるようになった. がん, 高血圧, 糖尿病などでも遺伝性のものがあり, この場合は原因となる単一の遺伝子が知られている. これに対し遺伝性でないものでは, 複数の遺伝子の変異が少しずつ発症の可能性を上げる形で寄与していると考えられている. したがって, これらの遺伝子は疾患の原因遺伝子ではなく, 感受性遺伝子と呼ばれる. また, 多くの疾患は複数の感受性遺伝子が関係していることから多因子疾患とも呼ばれる.

ヒトの遺伝子は様々な細胞において, 様々なタイミングでタンパク質を作り, そのタンパク質が他の遺伝子の作るタンパク質と非常に複雑なネットワークを形成している. このネットワークが環境に敏感に反応し変化することで生体機能が維持されている. 過大な環境要因が加わり, いくつかの遺伝子の機能に偏りがあるとネットワークの一部に不具合が生じ, 疾患が生ずると考えられている (**図1**). 環境要因がない場合には, 遺伝子の機能に偏りがあっても発症しないし, 環境次第ではその偏りが有利に働く場合もある. 現在患者数が急激に上昇している糖尿病や喘息・アレルギーなどの疾患では, 現代の人類が経験している環境の変化によって, それまで生存に有利であった遺伝子の機能の偏りが疾患感受性を増す方向に働いていると考えられている.

ありふれた疾患を予防し治療していくためには, 発症・進展に関わる感受性遺伝子と疾患遺伝子やその他の遺伝子が作るタンパク質が織りなすネットワークが解明される必要がある. この解明には大変な労力を必要とし, ヒトゲノムプロジェクトによって, ヒトゲノムのDNA配列が明らかになる前は, 可能であるとは考えられなかった. ヒトゲノム配列は究極の遺伝地図であると同時に, 全ての遺伝子

Presented by Medical*Online

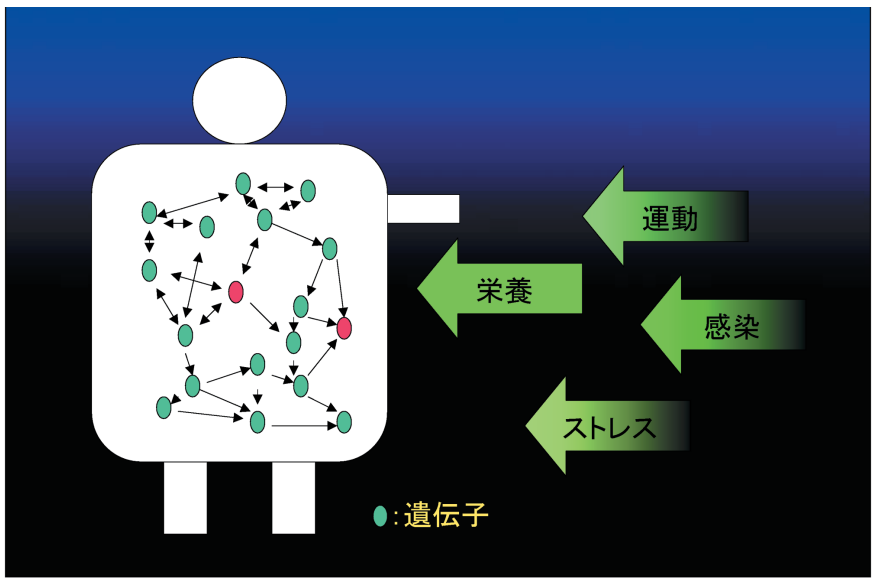

図 1　ゲノム医科学の考え方

表　ゲノムと推定遺伝子数

	遺伝子数	ゲノムサイズ(塩基対)	遺伝子当たりの塩基数
＜原核生物＞			
マイコプラズマ	470	580,000	1,234
らん藻	3,168	3,570,000	1,127
枯草菌	4,100	4,200,000	1,024
大腸菌	4,288	4,640,000	1,082
シュウドモナス菌	5,570	6,300,000	1,131
＜真核生物＞			
分裂酵母	4,944	12,500,000	2,528
出芽酵母	5,805	13,000,000	2,239
線虫	19,000	100,000,000	5,263
ショウジョウバエ	13,800	120,000,000	8,696
シロイヌナズナ	25,000	125,000,000	5,000
ヒト	31,200	3,000,000,000	96,154

を含んでいる. したがって, ヒトゲノム配列が明らかになれば, 疾患感受性遺伝子の同定にも, 遺伝子とタンパク質が織りなすネットワークの解明にもつながっていくと考えられる. そもそもゲノムプロジェクトは, そのような理由から医学に大きな進歩もたらすものとして計画された経緯がある. すなわち, ヒトゲノムに医学研究の基盤を置くことにより, 現代社会が直面する多くの疾患についてその遺伝要因が明らかにされ, 発症メカニズムが解析されると期待されているのである.

ゲノム配列の問題点とヒトcDNAプロジェクト

本年 4 月 14 日にヒトゲノム配列の完全解読が宣言された. いよいよ本格的にヒトゲノムに基盤を置いた医学研究の時代へ突入したわけである. しかしながら, ここにきてヒトゲノム配列だけでは基盤として不十分であることが分かってきた. **表**にこれまで分かったいくつかの生物のゲノムサイズと遺伝子数の関係を示す. この表から明らかなように, ヒトゲノムの特徴は遺伝子数に比べてゲノムサイズが著しく大きいということである. ヒト遺伝子はヒトゲノムの約5%程度と考えられ, タンパク質をコードしている部分は1.5%程度である. ヒトでは線虫やショウジョウバエに比べて, 遺伝子がゲノムの広い範囲に散らばっており, また1つの遺伝子も短い多数のエクソンと長いイントロンからなっていることを反映している. このために, ヒトでは従来他の生物用に使用されてきたコンピュータプログラムを用いて, 遺伝子とそれがコードするタンパク質を

Presented by Medical*Online

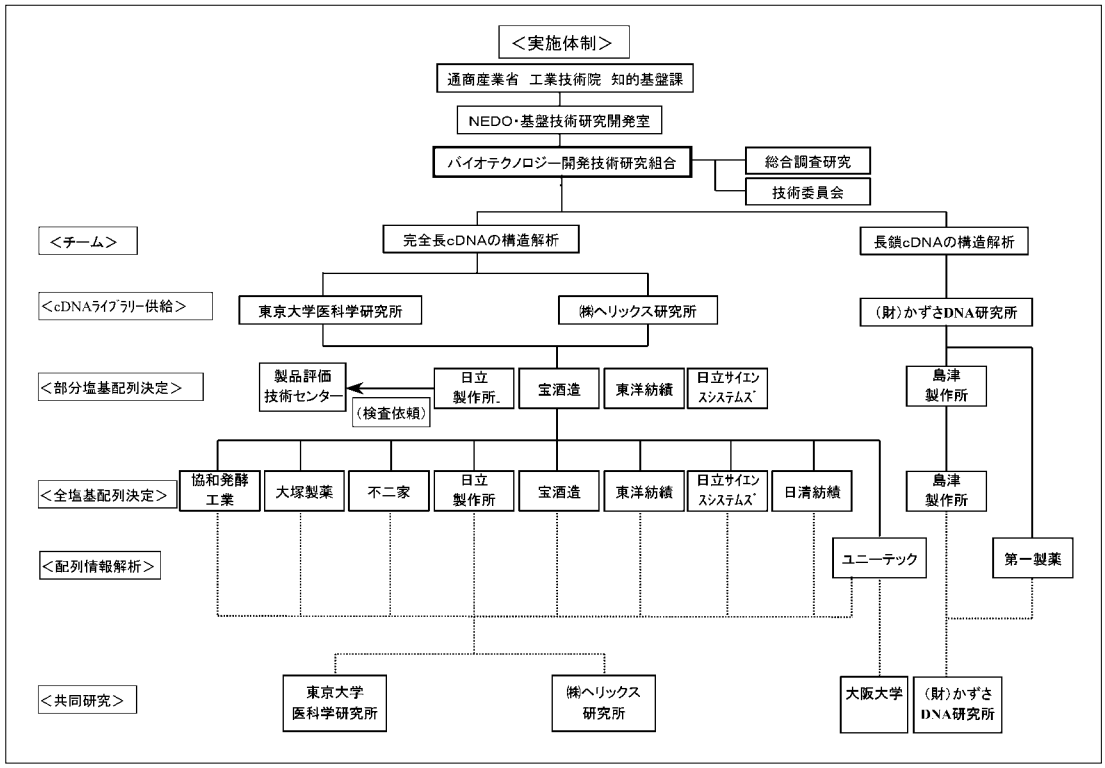

図2 完全長 cDNA 構造解析

正確に予想することが困難となっている. こうした状況はゲノムを基盤に医学研究を展開しようとする場合, はなはだ都合が悪い. そこでにわかに注目されるようになったのが完全長の cDNA である.

完全長 cDNA は mRNA の完全なコピーである. 完全長cDNAが存在すれば, ゲノム上の遺伝子の位置やそのエクソン・イントロン構造がわかり, それがコードするタンパク質が正確に分かることになる. また, 完全長cDNAは直接実験に使用可能で, 研究を行う上で有用な材料・手段になる. したがって, ゲノム配列が得られ, それが今後の医学研究に使うには不十分のものであることが明らかになると同時に, 完全長cDNAに注目が集まるようになった. ただ, 問題はmRNAの完全なコピーである「完全長cDNA」のクローニングの技術的困難さである. われわれが開発したmRNAの5'端に存在するキャップ構造をターゲットにした完全長cDNA選択法であるオリゴキャップ法は, この技術的困難さを解決し, 大規模なcDNAコレクション・配列決定プロジェクトに道を開いた. それが後述する経産省支援で行われたFLJプロジェクトである. このような大規模cDNAプロジェクトが日本で立ち上がったことで, 日本は世界のcDNA研究の中で確固とした優位を築いている. これは今後の医学・生物学研究を展開していく上で有利な基盤を確立したということができよう.

FLJプロジェクト

FLJプロジェクトはヒトの完全長cDNAのコレクションを目標に, 経産省・NEDOの支援で行われたプロジェクトである. **図2**にその体制図を示す.

ここに示すように, 本プロジェクトは東京大学医科学研究所, かずさDNA研究所, 半官半民で設立されたヘリックス研究所, 製品評価技術センター, それに日立をはじめ, 民間会社11社が一体となって行ったプロジェクトである. その中心は完全長cDNAにあり, まず完全長cDNAライブラリーを作製し, そのライブラリーからcDNAクローンを多数分離し, 5'端の配列を決定する. この5'端配列を元に, 新規で完全長らしいcDNAクローンを選び, その全長配列を行う. 作製したcDNAライブラリーは約130種類, 総数で約150万クローンをcDNAライブラリーから分離し, その5'端配列を決定した. この中から選ばれた3万クローンを全長解析した. 配列決定をした3万個のcDNAクローンのうち, 約2万2千個の配列データは国際公共配列データベースに公開している. 残りの約8千個については, 現在特許申請準備中であり, 申請が終われば順次公開の予定である. 3万個のcDNA配列の情報解析については, 現在いろいろな角度から解析を行っている. 大ざっぱに言うと, 1) ゲノム上の位置, 2) 既知遺伝子との相同性, 3) タンパク質機能モチーフの有無, 4) 膜局在シグナル・膜貫通部位の有無, 5) cDNA ライブラリーと5'

Presented by Medical*Online

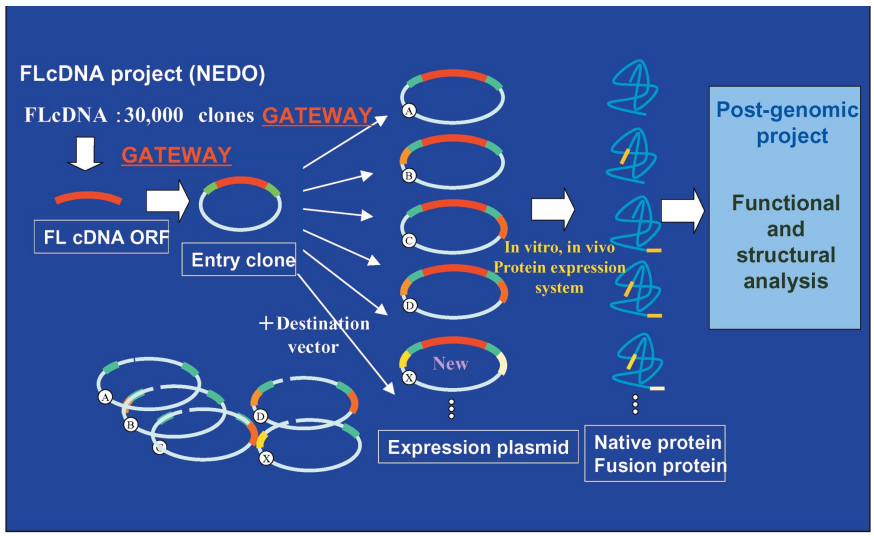

図3　High-throughput production of expression plasmid of FL cDNA

端配列情報を利用した遺伝子発現情報，などである．これらの情報を組み合わせて，公開したものの中でも，約8千クローンには何らかの形で特許申請がなされている．

ヒトの完全長cDNAのコレクションを作ろうという動きは日本が先行しているものの，各国でも始まっている．アメリカはNIHの中の国立癌研究所（NCI）が中心になり，年間50億円の予算でヒトとマウスのcDNAコレクションを行っている．ヨーロッパではドイツの癌研究所が規模は小さいものの，ヒトのcDNAコレクションを進めている．中国も上海のゲノムセンターがcDNAの解析を進めている．とはいうものの，現在データベース上で公開されているヒト全長cDNA配列データ（必ずしも完全ではない）は，日本：約23,000，ドイツ：約5,600，米国：約14,000，中国：約1000などで，日本のものが半数以上を占め，未だリードを保っている．マウスでも理研の林崎らのグループが圧倒的なcDNAクローンを持ち，イネについても完全長cDNAでは日本の独り舞台である．日本は完全長cDNA大国ということができる．

cDNAクローンからの展開

FLJプロジェクトをはじめ、世界中で行われているcDNAプロジェクトのおかげで多量のcDNAクローンが利用可能になり，その配列とゲノム上の位置が明らかになってきている．まさに機能未知の遺伝子が数万個も見つかってきたと言える状況である．今後はこのような多数の遺伝子を基盤にして，それらと疾患との関係について解析していかなくてならない．そこで重要になるのは，「網羅的」な遺伝子解析である．これは従来のように遺伝子を1つ固定して，それについて様々な解析を行い，疾患との関係を検討するのではなく，解析法を1つに固定して，多数の遺伝子について，1つの解析をいっぺんに行うやり方である．網羅的解析の大きな欠点は，得られる情報が一つ一つの遺伝子については断片的であることである．網羅的解析として，何万という遺伝子のmRNA量を同時に測定するDNAチップ・マイクロアレイ解析が有名だが，遺伝子の発現レベルが分かっただけでは，すぐに疾患に結びつくものではない．二次元電気泳動と質量分析機を使用したプロテオーム解析でも同様である．

しかし，網羅的解析は1種類の解析に集中するので効率が良く，データ取得速度が速い．しかも，思いがけない遺伝子と疾患の関係を明らかにすることがある．特に，効率の良さが大きな違いを生み出すことを示したのがヒトゲノム計画である．ヒトゲノム計画では配列決定に集中し，従来は不可能とみられていた全ヒトゲノム配列決定を短時間で完成させた．この成功が網羅的方法論の可能性を広く知らせることとなったのである．

網羅的解析を行う上で重要なことは，そのための技術基盤が存在することである．実際，cDNAコレクションで得られた配列データと遺伝子クローンは網羅的遺伝子解析を進める上での中核的な技術基盤と考えることができる．マイクロアレイやDNAチップは，このようなcDNA配列情報やcDNAクローンが利用できて初めて可能になったものである．現在注目の的になっているプロテオーム解析でも同様のことが言える．

網羅的トランスフェクション解析

現在われわれは，経産省の支援を受けて産業技術総合研究所生物情報解析センターにてcDNAクローンを用いた網

Presented by Medical*Online

図4 トランスフェクション・ロボット

羅的解析の様々な技術基盤を作り上げてきている．まず，cDNAクローンを網羅的に取り扱うことができる多目的発現系の基盤を確立することが重要である．このような目的のために開発されたGatewayクローニングシステムは効率性や汎用性に優れ，その有効性が示されている．現在，1万種類以上のcDNAのタンパク質コード領域がGatewayクローニングシステムに導入された．これらのクローンは，後述のようにいろいろな種類の細胞に対する，いろいろな目的での大量トランスフェクション実験に利用できることはもとより，*in vivo*/*in vitro*でのタンパク質産生など，その利用範囲は極めて広く，網羅的な機能解析プロジェクトの基盤となっている（**図3**）．

さて，第一の網羅的機能解析プロジェクトは高速トランスフェクション・プロジェクトである．このために，1日1万種類のcDNAクローンを各種の細胞にトランスフェクションするロボットシステムを作った（**図4**）．このシステム用いて各種細胞にcDNAを注入したのち，細胞の形状変化を定量的に測定する．この細胞の変化量と遺伝子の類似性データや発現頻度情報等のデータを総合的に解析することによって，注入された遺伝子の機能を推定することができる．

従来，細胞トランスフェクションは，各ステップがほとんど人手により行われていたため，細胞の形状変化の解析に有意なデータを取得するために多くの労力と日時を要しており，多数遺伝子を並列して同時にトランスフェクションをし，機能の推定を行うというのは事実上不可能であった．本システムでは，遺伝子を注入した細胞の変化を顕微鏡から画像データとして取り込み，最新の画像情報処理解析技術によってその形状の変化を検出し，遺伝子機能の解析のためのデータ取得・解析を高速に行うことにより網羅的形状解析が可能となる．現在，このシステムとGFP融合型のタンパク質を発現できるGATEWAYクローンを用い，網羅的なタンパク質の細胞内局在解析を行っている．

プロテオーム解析

ヒトゲノム配列や完全長cDNA塩基配列情報の蓄積の進行により，包括的なタンパク機能解析が可能となりつつある．多くのタンパク質はその機能を発揮する際，他のタンパク質と相互作用をして情報を伝達していく．このようにタンパク質-タンパク質相互作用はタンパク質の機能の基本であるが，相互作用を検出する方法としては，現在，酵母菌ツーバイブリッドの系がよく使われている．しかしながら，この方法には人工的で特殊な環境での反応である（細胞質で発現したタンパク質が核内に輸送される必要がある）ことや，検出できるのは2つのタンパク質分子間の結合に限られることなどの問題点が指摘されている．したがって，生細胞内の生物学的活性を保ったタンパク質複合体を同定するには別の新たな方法が必要である．

そこで，われわれはMycタグ，Hisタグのようなタグを持ったタンパク質をGATEWAYクローンを用いて細胞内で発現し，細胞内で相互作用して結合しているタンパク質を複合体の形で同定することにした（**図5**）．このとき，タンパク質の同定には質量分析計を用いる高速同定システムを用いる．特に，同定感度と精度を上げるために都立大の磯部らが開発したナノLCシステムを用い，タンパク質の同定を行っている．ナノLCシステムは文字通り微細な

Presented by Medical*Online

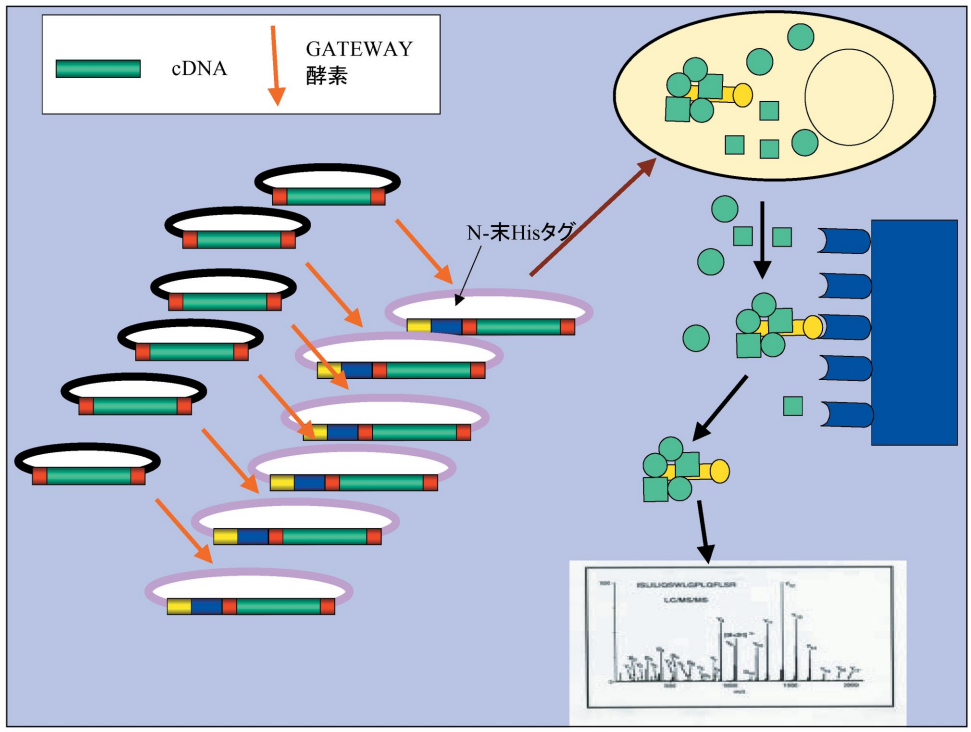

図5　タンパク質相互作用の網羅的解析

キャピラリ型のLCカラムを用いて，タンパク複合体を直接トリプシンなどの酵素で消化して得られるペプチドを分けた上で，質量分析計に導入するシステムである．従来の2次元電気泳動を用いる方法に比べてステップが少ないためロスが無く，検出できるタンパク質の数も多く，検出感度も高い．

この方法を用いることで，従来から相互作用が知られているタンパク質群の他に，様々なタンパク質が結合することが分かってきた．その多くは精製過程でのコンタミか，タンパク質をある意味非生理的レベルで発現したために起こったと考えられるが，いくつかの物は正しく，遺伝子機能の新しい局面を示す物であると考えられる．このような情報を現在蓄積されつつあるヒトゲノム配列や遺伝子発現情報と関連付けることで，さらに結合することの意味の推定が期待できる．実際，タンパク質の相互作用解析を遺伝性疾患の原因遺伝子であることの分かっている物について網羅的に行うことで，遺伝性疾患の発症機序に迫ることができると考えている．単一の遺伝子が原因で起こる遺伝性疾患でも，多くの症状は対応する多因子病と共通しており，発生機序もある程度の重複があるものと考えられる．したがって，遺伝性疾患の原因遺伝子と相互作用するタンパク質を網羅的に解析することで，それに対応する多因子疾患でも重要な働きをしている遺伝子を捕まえる可能性も高いのである．

プロテイン・ファクトリー

タンパク質の機能を解析する場合に，タンパク質を大量に発現させ精製した上で機能解析をする場合がある．抗体の作製，酵素活性の測定，3次元構造の解析などである．しかし，タンパク質の発現系はいろいろな種類の物があり，タンパク質の性質に応じ，解析したい内容に応じて発現系を選択する必要がある．この部分はコストも時間もかかる部分であり，従来のタンパク質研究の隘路になっていた．GATEWAYクローンの導入で様々なタンパク質発現系に対応したベクター間を，cDNAを容易に動かすことができるようになった（**図6**）．このような状況を踏まえて，どのタンパク質がどのタンパク質発現系で発現しやすいか，パイロット的に検討するプロジェクトが動き始めている．この検討結果は，データベース化して将来利用可能にする予定である．タンパク質3000プロジェクトをはじめ，タンパク質の3次元構造を決め，それを医薬品などの開発に役立てようというプロジェクトが国内外で始まっており，これはそのための貴重な基礎情報を与えるであろう．

このプロジェクトには，多数の企業が自分が得意なタンパク質発現系を引っ下げて参加しており，このようなオールジャパンとしての取り組みも近頃のプロジェクトとしては珍しい．タンパク質の発現系は*in-vivo*系と*in-vitro*系に分けられる．*In-vivo*系は，大腸菌，枯草菌などの細菌類を使うもの，酵母などの真核微生物を使うもの，昆虫細胞や

Presented by Medical*Online

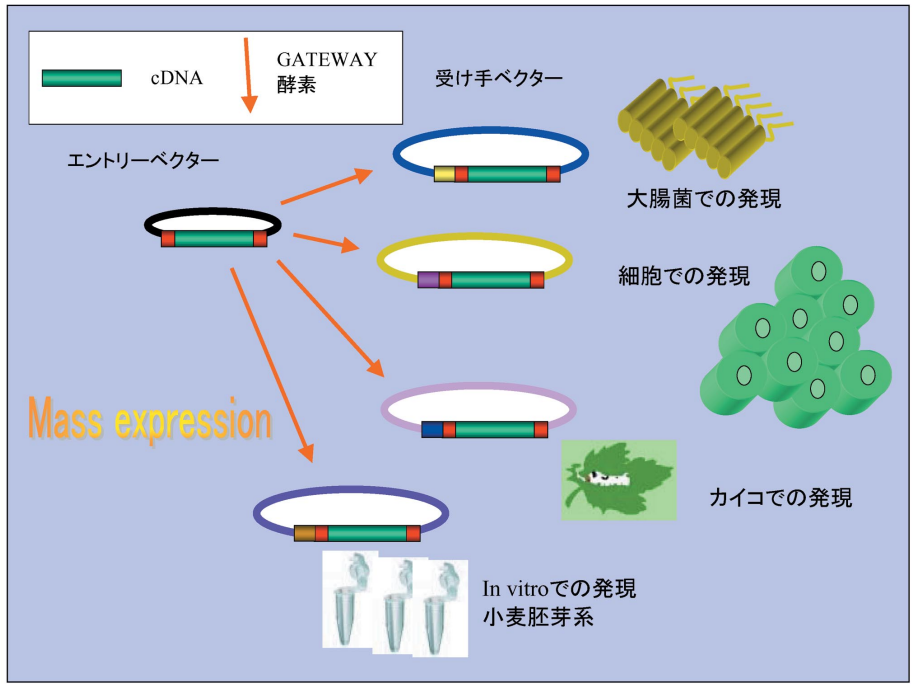

図6 プロテイン・ファクトリー・プロジェクト

幼虫を使うもの, 培養細胞を使うものなどがある. *In-vitro* 系には, ウサギ赤芽球抽出液や大腸菌抽出液, コムギ胚芽抽出液などが用いられ, 近頃進歩が著しい. 特に, 多数のcDNAクローンを用いて網羅的にタンパク質を発現するには, ロボットなどと相性の良い *in-vitro* 系が第一選択になっている.

現在までのところ, *in-vitro* 系では愛媛大学の遠藤らが開発した改良コムギ胚芽抽出液が, 発現効率で最も良い成績を上げている. 膜タンパク質など, 原理的に苦手のはずのタンパク質も結構発現している. ただし, 膜タンパク質などでは本来の形を保っているかどうかは不明である. 発現量もある程度見込めるが, 現在はコストが高いので用途が限られる. 大腸菌の *in-vitro* 系では, この系で発現しやすいタンパク質と融合タンパク質を作ると発現効率が上がるようである. *In-vivo* 系では, 片倉工業のカイコの系が面白い. cDNAを一つ一つウイルスベクターに入れてウイルスを作らなければならないところが大変だが, GATEWAY システムとの併用もあり, その部分が少し易しくなっている. ウイルスが一旦できると培養昆虫細胞だけでなく, 虫体そのものにも注射して, 大量にタンパク質を発現させることができる. もともと, 膜タンパク質には良い系であり, 大量に発現させた膜タンパク質をどのように精製するかというようなところが今後の課題である. これからは, このプロテイン・ファクトリー・プロジェクトの結果を使い, プロテインチップの作製等を試み, 網羅的な酵素活性解析や網羅的リガンド探査などにつなげていく予定である.

おわりに

ヒトゲノム配列が決定されポストゲノムの時代に入ってきた. ポストゲノム時代の疾患研究では, モデル生物の網羅的遺伝子破壊に対応する人類遺伝学・集団遺伝学的解析と, ここで紹介したcDNAを基盤とした網羅的遺伝子機能解析が車の両輪となる. このようなゲノムを基盤とする疾患研究と, 従来のより生物学的な疾患研究を上手く組み合わせることで, 様々な疾患に関わる遺伝子の同定が進み, 発症メカニズムの解明ができると考えられる.

そして, 「なぜ治療効果が上がるのかはわからないが, 経験的に効果があるから」といった医療から, 疾患の成因, メカニズムを見極めた上でその原因となっているものを取り除く, あるいは発症経路を阻害するといった理詰めの医療への動きもさらに進むと考えられる. また, ゲノム配列の個人差についての情報も1塩基多型 (SNPs) を中心に急速に蓄積されつつあり, 個人個人の病気の質を見極め, 一人一人にあった個人化医療の実践に向けた基盤も整いつつある. 医療におけるこうした劇的な進歩に, ヒト完全長cDNAがどのように貢献しているか, ご理解いただけたことと考える.

Presented by Medical*Online