生命情報学とは，情報学の考え方により生命科学の諸問題にアプローチする学問である．ゲノムやタンパク質等の生命データ処理に内在する情報学的問題を探して解き，情報学コミュニティで発表する立場もある．しかし一歩踏み出して，生命科学の謎を情報学として定式化して解く視点の方が，生命科学者から理解され歓迎される．そのような例を本稿では与える．近年，DNA 解読装置の革命的進歩（ゲノム情報ビッグバン）により，ムーアの法則を凌ぐ約 8 カ月あたり 2 倍の速度で解読速度が上昇している．2003 年に解読が宣言されたヒトゲノムには約 500 万ドルの費用がかかったが，2010 年現在では数日で数万ドル程度，2 ～ 3 年後には 1 時間以内に 1,000 ドル以下で解読が可能になると言われている．この技術的進歩により解明が可能になるであろう医学および生物学の問題を紹介し，情報学的思考が必要な場面を例示する．

生命情報学

ゲノム情報ビッグバンの進展

道しるべ

森下真一 （東京大学）

DNA 解読のこれまで

生命情報学（バイオインフォマティクス）が世間でも比較的知られるようになったのは 2000 年以降に報告されたヒトゲノム解読のころであろう．解読は 1986 年に計画されたものの，当初は遅々とした歩みであった．ヒトゲノムは全体で約 30 億の塩基対から構成され，22 本の常染色体と 2 本の性染色体を持つ．各染色体を端から端まで一気に読むことができる DNA 解読装置が存在すればよいが，当時から現在に至るまでそのような装置は出現していない．たとえば 1990 年後半から 2004 年ごろまでに市場に普及していた DNA 解読装置は，約 500 塩基を 5% 程度のエラー率で解読するのが精一杯であった．現在でも，1,000 塩基以上の DNA 断片を安定的に解読できる装置はほとんどない．

DNA を解読するには，DNA を細かく断片化し，電気泳動させ，泳動した距離から塩基の並び順を間接的に推定するサンガー法[1)]が主として使われてきた．泳動した距離の分解能には限界があるため解読

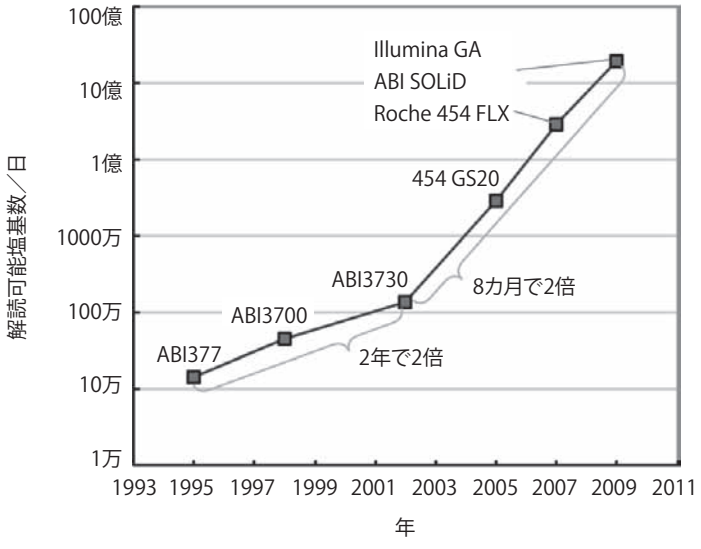

図 -1　DNA 解読装置の解読量の変遷

長は 500 ～ 1,000 塩基程度になる．さらに 1 本の断片の解読に 2 時間程度かかるため，塩基解読量を増やすために同時に並列に読むようになった．2002 年ころには 384 本同時に解読する装置 ABI3730 が普及し，1 日約 200 万塩基を解読できるようになる．それでも 30 億塩基対の解読は長い道のりである．多数の DNA 解読装置がヒトや多様な生物種のゲノム解読のために普及した．2003 年にヒトゲノム解読が宣言され，2004 年には解読終了論文が Nature に発表される [2]．共著者が 2,000 名を超えるビックサイエンスの論文で，コストは約 500 万ドルかかったと言われる．このあと研究コミュニティは過剰な盛り上がりから落ち着きを取り戻す．解読されたヒトゲノムは，ヒトゲノムの個人差の分析，遺伝子機能の分析等へ多様に応用されている．

ゲノム情報ビッグバン

ヒトゲノムを解読した後は，哺乳類や脊椎動物のゲノムの解読が進んだが，ヒトゲノム解読ほどは盛り上がらなかったと感じる．我が国におけるゲノム計画への熱は徐々に冷めていった．一方，アメリカでは熱が冷めるどころか，次の大規模なプロジェクトである 1,000 ドルヒトゲノム計画を 2004 年に発表する．読んで字のごとく，1,000 ドルでヒトゲノムを解読する技術開発を支援する研究予算であった．この計画が発表されたとき，日本では一部の研究者は注目していたものの，夢物語を聞かされているような感じでもあった．しかしこれが本格的な計画であることを毎年実感するようになる（**図 -1**）．

サンガー法を使わない計測技術 pyrosequencing[3] により，長さ 100 塩基前後の配列を約 20 万個並列に 4 時間で解読する Roche 社の 454 GS20 シーケンサが 2005 年に登場した [4]．20 万という並列度に肝を潰したが，長さが 100 塩基程度であり，サンガー法の 500 塩基よりは遥かに短いため，応用は限定的と考えられていた．しかし改善は目覚ましく，現在ではサンガー法に匹敵する 500 塩基前後の配列を約 100 万並列で解読できるようになっている．

2006 年後半からは，長さ 25 塩基の塩基を約 5,000 万個も並列で解読できる Illumina 社の Genome Analyzer（通称 Solexa）が普及する [5]．長さが 25 塩基と短いものの，1 回の実験（3 日程度）で 10 億を超える塩基を収集できるのは革命的であった．それ以前までは，あまりにコストが膨大になるので観測がためらわれていた対象（後述）に手が届くようになった．年々改善が進み，2010 年には長さ 200 塩基の DNA 断片を約 10 億個並列で解読する HiSeq 2000 が発表された．8 日間で約 2,000 億塩基が収集可能と言われている．以上の方式は短い DNA 配列を超並列で解読することにより収集する塩基量を確保している．ボトルネックは 1 塩基の認識速度であり，Illumina 社の場合，約 1 時間を要する．

一方，自然界では驚くほど高速に塩基を識別してゆく．DNA の複製時に 2 本鎖 DNA は 1 本鎖にほどかれ，DNA ポリメラーゼが 1 本鎖上の塩基を解読しながらコピーする．ヒトゲノムの場合，秒速 60 ～ 90 個もの塩基を複製する．約 1,000 ～ 10,000 カ所で同時に複製が起こり，約 8 時間でヒトゲノム全体をコピーする．超高速かつ長く DNA を読むために，この仕組みを使うことができな

いか？　2008 年 2 月にシリコンバレーの Pacific Biosciences 社はこの考えを実現し注目を浴びた [6]．1 本鎖の DNA 断片が，DNA ポリメラーゼにより 1 塩基ずつ複製されてゆく過程で，塩基ごとに異なる色を発光するように工夫し，光を検知する．DNA1 分子を検知でき，しかも秒速 1 ～ 2 個の塩基を解読できる．1 塩基の読み取りに 1 時間を要する Illumina 社方式に比べて速度は数千倍向上する．さらに塩基解読長も非常に長く，1,000 ～ 3,000 塩基程度は解読できるようである [7]．ただし並列化は今後の課題であり，2010 年の段階では約 8 万並列にとどまっている．そのため塩基出力能力は Illumina 社の新型解読装置 HiSeq 2000 に及ばないと予測される．今後，並列度を向上し，塩基読取速度が自然界の秒速 60 ～ 90 個に近づけば，他の方式を凌駕するであろう．2014 年に秒速 1 億塩基を目指しているそうである．

Pacific Biosciences 社方式の他の特長は，微量の DNA サンプルで解読できそうな点にある．より詳しくは，1pg（10^{-12}g）がおおよそ 10 億塩基対であり [8]，ヒトゲノム 2 倍体で約 6pg になる．しかし現在の DNA 解読装置では DNA サンプル量が 1 ～ 10ug（10^{-6}g）ぐらい必要になる．細胞数で換算すると 10 万個程度が目安になる．これだけ大量の細胞をあつめる作業は，微量の細胞（自然界の幹細胞，受精後の 2，4，8 細胞期の細胞など）では困難である．DNA サンプル量が 1ng（10^{-9}g）程度で済むような DNA 解読装置が登場すると，観測対象はとても広くなる．1 分子計測技術はその点で重要である．1 分子計測により高速に解読する方式は，他社からも提案されている．2010 年には，塩基を発光させる代わりに PH の変化を検知する方式が Ion Torrent 社から発表され，より低コストで DNA 解読を実現できる技術として注目されている．

図 -1 にこれまでに普及した解読装置 1 台あたりの塩基解読量を時系列に沿って表示した．2002 年以降は 8 カ月で 2 倍の速度であり，当面はこの勢いは続くと考えられる．DNA 解読の進歩は激流の中にある．

医学への展開

家族性癌など遺伝の要素の強いさまざまな病気や，薬の効果の個人差，酒の上戸・下戸（アルコール分解能力の個人差），集中力の個人差などはゲノムの個体差に由来すると考えられている．こうした個体差はさまざまであり，遺伝子の塩基配列のうち，1 つの塩基が別の塩基に置換されていたり，一部の塩基配列が欠落／挿入されたり，ゲノム中にコードされる遺伝子のコピー数の違い，短い塩基配列の重複数の違いといったことから生まれる．

最初に解析されたヒトゲノムは，不特定多数から無作為抽出された匿名の人物のサンプルであった．2008 年以降，個人のヒトゲノムがいくつか公表されている．DNA の 2 重らせん構造を発見した Watson 博士 [9]，ヒトゲノム解読で著名な Ventor 博士，匿名のアジア人 [10] および韓国人 [11] 等のゲノムである．分析の結果，予想以上に細かい違いが多く，人種間の格差は大きい．一方，人種内の多様性は小さくなる傾向にあると考えられるので，日本人の標準的なゲノムの整備が進んでいる．また，米英中を中心とする研究機関が共同で「1,000 人ゲノムプロジェクト」を進めている．匿名性を保証しながら，1,000 人のゲノムを解析することで，塩基配列と個体の多様性の関係について，さまざまな知見が得られると期待されている．

ヒトゲノムの解読は病気に関連する遺伝子を探す場面できわめて有効である．どのように探すかと言えば，ヒトゲノムに数百万個存在すると考えられる 1 塩基の変化（SNP : Single Nucleotide Polymorphism）を目印に疾患感受性遺伝子を探索することが常套手段となっている [12]．SNP は集団内で高頻繁（5% 以上）に観測される SNP から，稀に起こる低頻度（0.1 ～ 1%）の SNP まである．病気との関連が深い低頻度の SNP ほど，病気の診断に有効であることが分かってきた．しかし単一の低頻度 SNP が，特定の病気に罹患する集団全体で共有されていることは少なく，単一の低頻度 SNP で網羅的な診断をすることは難しい．

そこで複数の低頻度 SNP を多数集めて，そのどれかを保有するか否かで診断する方式（multiple rare variants）が最近注目され，有効性が吟味されている [13]．このためには，低頻度 SNP を多数プロファイリングすることが必要であるが，コストが高くつく．たとえば約 1% の頻度の SNP を見つけるには，少なくとも 100 名以上のサンプルを使い頻度を狭い信頼区間で推定することが望ましい．しかしサンプル数が多いほどゲノム解読は高コストとなるため敬遠されてきた．そのため従来は，サンプルが少なくても推定できる高頻度の SNP がもっぱら収集され，人類遺伝学で愛用されてきた．幸いゲノム情報ビッグバンの進展は，低頻度の SNP を低コストでプロファイリングするための良い見通しを与えている [14]．しかし解決しなければならない問題も残っている．DNA 解読にはエラーがつきものなので，読取エラーを自然界に存在する SNP と誤って判断することは避けなければならない．低頻度 SNP を収集するコストを最適化する問題を解決できれば生命科学への貢献も大きいであろう．ただし実験コストを最適化するためには，実験の詳細にも精通する必要がある．ある種の仮定をして問題を定式化し，アルゴリズムを最適化したにもかかわらず，より簡潔な実験方法を思いついて，精度を格段に上げることに成功する場合もある．したがって情報学と計測技術の双方を鑑みた考察が大事である．

個人のヒトゲノム分析は慎重に進めなければならない．具体的には「遺伝学的検査に関するガイドライン」と「ヒトゲノム遺伝子解析研究に関する倫理指針」を遵守しなければならない．DNA サンプルの提供者が遺伝子情報により差別を決して受けない配慮が必要である．またサンプルを提供していただく際には，研究の内容を説明し同意を得なければならない（Informed Consent と呼ぶ）．これらは医学部で遵守する規則である．情報を分析する立場の研究者は個人遺伝学的情報の機密性を保証しなければならない．データサーバ室への生体認証ゲートの設置，サーバのインターネットからの隔離，データ転送の暗号化などの対策が通常は求められている．今後は暗号化されたデータベースに対する高速な問合せを実現することを検討し，機密性を保証しつつ，分析の利便性を高める工夫が必要であろう．

生命科学分野への展開

医学以外の応用も広がっている．たとえば，経済動物（ニワトリ，ウシ，豚，など）や経済植物（イネ，小麦，キュウリ，キャベツ，イモ，ぶどう，など）のゲノム解析は盛んである．収穫の多い作物に特徴的なゲノム配列を探索することは，品種改良へとつながる．また，冷害や乾燥，害虫などに強い作物のゲノムを解読して，悪環境に強い個体の塩基配列の特徴を検出し，品種改良を行うことで，寒冷地や乾燥地帯など，これまで作物が育ちにくかった地域を農地とすることができるかもしれない．

エネルギーや環境問題へのアプローチもある．トウモロコシやサトウキビからエタノールを作るバイオ燃料は，次世代エネルギーとして期待されており，実際ブラジルでは普及している．ただし，原料のトウモロコシの高騰を招くなどの問題を孕んでいる．そこで代替エネルギー源となる生物が探されている．たとえば植物から燃料を醸造するには，酵母や菌類を介在させるが，エネルギー変換を行う酵素遺伝子をさまざまな生物のゲノムの中に探すことで，廃材やサボテンなど穀物ではない植物から，効率よく燃料を醸造する技術が生まれる可能性がある．そのため米国エネルギー省は NIH（National Institutes of Health）と並んで，ゲノム解読に力を入れている．

環境問題では，たとえば重金属により汚染させた土壌は開発途上国で深刻な問題となっている．このような汚染土壌を化学的処理で土壌改良するには高いコストがかかるそうである．最近，筆者の周辺にいる大学院生から，汚染土壌で生育できる植物は少ないが，好んで生息する苔類があるという話を聞いた．彼は風の谷のナウシカの腐海に例えていた．これらの苔類は重金属に対する耐性を持ち，重金属を体内に蓄積する苔もあり，重金属の回収に役立つ可能性がある．何らかの生命システムを獲得している

はずで，彼の話をきっかけにゲノム解読を開始している．

以上，医学，食糧，エネルギー，環境問題にわたって，ゲノム解読がどのように応用されているか概観した．ゲノム情報ビッグバンは，応用だけでなく基礎生物学の未解明な問題にも，明るい兆しをもたらしている．DNA の塩基配列は 1 次元の文字列であるが，実際にはヒストン 8 量体という蛋白質多量体に 147 塩基対が巻き付いて，数珠（本当はヌクレオソームと呼ぶ）のような構造をとっている（**図 -2**）．数珠の間隔は 10 ～ 50 塩基対程度であるが，DNA 上でまちまちである．数珠の位置も固定されている場所と，動きやすい場所がある．数珠が外れた場所が遺伝子の読取を制御しており，生命現象の基本メカニズムである．逆に，数珠が外れないようにする仕組み（DNA メチル化）もあり，遺伝子の読取を封じている．さらに数珠はより高次なファイバ構造をとりながら DNA を 3 次元的にコンパクトに折りたたんでいるが，そのプロセスはほとんど分かっていない．生物学の大問題の 1 つである．

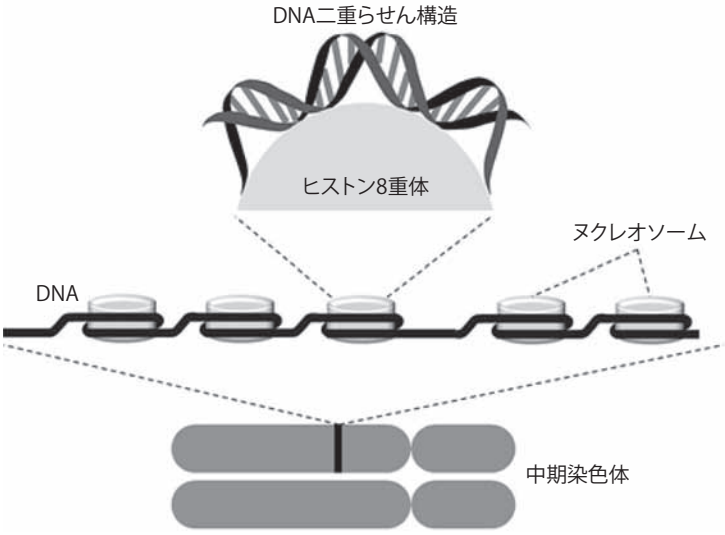

図 -2　ヌクレオソーム構造

これまでは，すべての数珠の位置を観測することは困難であった．数珠が平均して 200 塩基対に 1 つ出現するとしても，ヒトゲノム全体ではおおよそ 1,500 万個も存在するためである．ところが 1 回の実験で 2 億個もの DNA 断片を解読できる装置の出現により，数珠に巻き付いた DNA 断片だけを解読する実験方法を使って，数珠の位置を推定できるようになった（ただ実験は難しい[15]）．この結果，数多くのあらたな知見が得られた．遺伝子を読みとる位置の下流では数珠の位置が固定されやすい現象や[16]，数珠の位置が進化に影響を与える現象などが報告されてきている[17]．なお，数珠の位置を DNA 塩基配列だけから推定するアルゴリズムもいくつか提案されているが[18]，精度は高くなくあまり信頼されていない．DNA 塩基配列だけからは位置が決まらないのであろう．当面は観測情報を収集し，現象を説明する数理的モデルを構築してゆく研究が盛んになってゆくであろう．

もう 1 つ DNA の基本的な仕組みである DNA メチル化を紹介する．DNA は A，C，T，G の 4 つの塩基から構成されるので，情報科学的には 4 文字からなる文字列として単純化されて扱われる．しかし各塩基は高分子であり，化学的修飾を受け変性する．DNA メチル化とは，C（シトシン）塩基がメチル化を受ける現象である．いくつかのステップを経て数珠が外れないようにして遺伝子の読取を封じる．DNA メチル化はさまざまな生命現象の中で DNA の機能を変化させる．哺乳類では親のゲノムから DNA メチル化を遺伝的に継承するインプリンティングが特定の遺伝子の働きを抑える．受精直後の 2，4，8 細胞期には DNA メチル化がいったん解除され，再度メチル化される現象もある．癌細胞では DNA メチル化が癌抑制遺伝子の働きを抑えていると考えられる．

全 DNA にわたってどのような変化が起こるかを詳細に検討した研究はいまのところわずかである[19]．メチル化の状態を全 DNA にわたって観測するのが困難だったからである．しかしゲノム情報ビッグバンにより，メチル化を受けた C を特定する実験（これも難しいが）により，情報の収集が可

能になってきている．今後，さまざまな現象の理解が進むであろう．

必要なソフトウェア技術

DNA データ処理における基本的なソフトウェアに，DNA アセンブラとアラインメントがある．DNA 解読装置が 1 度に読める DNA 断片の長さは 50 ～ 1,000 塩基程度である．そのため，DNA 断片どうしがある程度重なり合うように，冗長に DNA を断片化して解読する．通常は元の DNA を 6 ～ 7 倍もカバーできるほどの量の DNA 断片を冗長に解読し，重なり合った部分を頼りに DNA 断片を繋げてゆく．ただし DNA 中に存在するコピーされた配列のために，DNA 断片を曖昧性のないまま繋げてゆくことは難しい．DNA を X 倍カバーすれば，理論的には X の指数関数に比例して，断片をつなげた長さの平均長は伸びてゆくはずである（Lander-Waterman の考察）．しかしコピーさせた配列がもたらす曖昧性により，伸張はコピー配列にさしかかったところで止めざるを得ない．コピーされた配列の曖昧性を避け DNA を長くする研究は重要である．哺乳類や脊椎動物ゲノムの解読が盛んであった 2007 ～ 08 年ごろまでは，サンガー法を使って解読された長さ 500 塩基前後の DNA 断片を対象にしていた．DNA 断片は長いほど，コピーされた配列が生む曖昧性の影響を受けにくくなり都合が良い．2008 年以降，ゲノム情報ビッグバン時代で主流になった長さ 50 以上の短い DNA 断片をあつかう研究が増えた．

哺乳類や脊椎動物の主要な生物種の DNA の解読はほぼ終わった．現在は，標準となる DNA が存在する種において，種内の多様性をプロファイリングすることが盛んで，特にヒトゲノムにおいて顕著である．標準となる DNA と，あたらしいサンプル間の「違い」に注目するため，サンプルから DNA 断片を解読して，標準となる DNA 中で対応する個所を探すアラインメントという情報処理をする．解読した DNA 断片には標準ゲノムとサンプル間の真の違いも存在するし，解読時の読取エラーも存在する．エラー除去のためには，多数の DNA 断片を解読して補正しなければならない．したがって，数 % の違いを許容して探すことが重要になる．

DNA は非常に長いので，あらかじめ前処理をして，アラインメントを高速化する．しばしば使われるのは，DNA 中の固定長の部分配列からハッシュ表を構成する方法と，接尾辞配列（suffix array）[20] を生成して Burrows-Wheeler 変換（BWT）[21] により配列を探索する方法である．ハッシュ表に比べ，接尾辞配列と BWT を使った方法は主記憶の消費が一般には少なく，任意長の問合せ配列を自然に処理できる．接尾辞配列の線形時間構築法は 2003 年に提案されたが [22], [23]，現実には計算時間がかかるため敬遠されてきた．その後改善が進み，2009 年に提案された induced sorting を利用した実装は，C プログラムで 100 行程度とシンプルでありながら，効率的に接尾辞配列を構築することが可能である [24]．学部等での講義題材としても適切であり，2 億塩基程度の DNA 配列であればノート PC でも扱うことができるので演習にも使える．現在最も使われるようになってきた BWA[25] や BWA-SW[26] というソフトウェアも suffix array と BWT により実装されている．

DNA アセンブリや DNA へのアラインメントは，いまだに研究の余地があるものの，おおよそ解決の見通しはついてきている．いま最も困難な問題となっているのは巨大ゲノムデータの高速処理であろう．2010 年度末には 1 日当たり 250 億塩基の解読能力のある装置が普及するであろう．ヒトゲノムの場合，両親に由来する 2 つの染色体の差を識別するために，1 人から約 1,000 億塩基を収集する．塩基情報に加えて，塩基の読取信頼度情報が付加されるので，1 塩基あたり 1 バイトの記憶領域が必要とすると 100GB ／人の情報が蓄えられる．このデータは多数の DNA 断片情報であり，ディスクから読み込み，ヒトゲノム上でアラインメントし，位置に関する情報を再びディスクへと書き込むが，書き込みデータ量は入力データ以上に大きくなる．多数

のCPUから大規模データの読み込みと書き込みが並列して実行されると，CPUとディスク間のデータ転送量がボトルネックとなる．ブロックサイズをある程度大きくし，ランダムアクセスは避け，シーケンシャルなデータ転送による実装を心がけると，200～500MB／秒程度の転送量を確保できそうである．現在の超並列計算機システムでは，どうしても主記憶，CPU，主記憶間のデータ転送の性能向上に優先度が置かれていて，主記憶とディスク間のバッファ管理に配慮したシステムが少ない．そのため，DNA情報処理には不向きなシステムが多い．2013～14年ごろには1日当たり1TBの情報量を生む解読装置が普及し，数万人分のヒトゲノム情報を処理する時代がくる可能性もある．ファイルアクセスが快適な並列計算機の構築が，DNA分析の鍵になるであろう．

世界の動向

ゲノム情報ビッグバン時代をもたらした解読技術は，米国の1,000ドルヒトゲノム計画を発端としている．アメリカおよびイギリス発の特許が製品の中に活かされ広がっており，日本は後れを取っている．DNA解読装置もアメリカが最も普及しており，日本と中国は台数では拮抗しているそうである．ただし，日本の場合は全国の研究機関に広く普及したのに対して，中国では北京ゲノム研究所に集中的に配備されている．装置を安定して稼働させるには熟練が必要である．集中的に配備した方が熟練したスタッフが高い稼働率で多数の機械を動作させてくれる．しかし技術が広まるわけではないであろう．日本の場合は，装置の稼働率を向上させるための情報交換が大切である．

中国の北京ゲノム研究所は急速に拡大している．ここ1年間で，パンダ，キュウリ，カイコ，腸内細菌などのDNA解析をNature等の雑誌に報告しており，世界屈指のDNA解読センターとなっている．2010年1月にIllumina社のHiSeq 2000を128台注文して大きな話題になった．1台1億円近くするとても高価な装置である．政府からの支援は少なく，民間から研究資金を獲得しているそうである．このようなオペレーションは勉強になるため，年1回，日中の若手DNA研究者を集めたワークショップを開催している．

おわりに

1日当たりのDNA解読量が8カ月で2倍になるというゲノム情報ビックバンについて概説した．DNA分析における情報学の役割は大きい．情報学的価値観だけから解けそうな部分問題を切り出して研究すれば，工学的喜びは得られるだろう．しかしサイエンスとしての面白さ，言い換えれば自然を観察し背景を洞察する喜びはなかなか得られないため，生物学者から共感を得ることが少ない．このジレンマに悩む人も多い．そこで，ゲノム情報ビッグバンが医学および生物学のどのような問題に対して解決の糸口を与えているかに力点を置いて解説した．情報科学的に解決しなければ問題も多いものの，本稿では少なめに抑えた．

1995年に最初の細菌DNAが解読され，それ以降にDNA分析は急速に膨らんできた比較的新しい研究分野である．生物学者だけでなく，数学者，物理学者がDNAサイエンスに惹かれて転向してきた場合も多い．一方，生物学者には，プログラミングを勉強して必要なソフトウェアを自分で作り，これからの生物学は情報学的思考も大切だと考えている人も多い．異なる分野を貪欲に学ぶことは大切である．東京大学でも，2003年に生命情報学を専門とする情報生命科学専攻が設置され，2007年には理学部に生物情報科学科が設置された．情報学と生物学双方の講義を開講しており，情報学的視点で生物学を研究できる人材を育んでいる．

道しるべ

参考文献

1) Sanger, F. and Coulson, A.R. : A Rapid Method for Determining Sequences in DNA by Primed Synthesis with DNA Polymerase, J Mol Biol, Vol.94, No.3, pp.441-448 (1975) .
2) International Human Genome Sequencing Consortium, Finishing the Euchromatic Sequence of the Human Genome, Nature, Vol.431, No.7011, pp.931-945 (2004) .
3) Ronaghi, M. : Pyrosequencing Sheds Light on DNA Sequencing, Genome Res, Vol.11, No.1 pp.3-11 (2001) .
4) Margulies, M., et al. : Genome Sequencing in Microfabricated High-density Picolitre Reactors, Nature, Vol.437, No.7057, pp.376-380 (2005) .
5) Bentley, D.R., et al. : Accurate Whole Human Genome Sequencing using Reversible Terminator Chemistry, Nature, Vol.456, No.7218, pp.53-59 (2008).
6) Korlach, J., et al. : Selective Aluminum Passivation for Targeted Immobilization of Single DNA Polymerase Molecules in Zero-mode Waveguide Nanostructures, Proc. Natl Acad Sci U S A, Vol.105, No.4, pp.1176-1181 (2008).
7) Eid, J., et al. : Real-time DNA Sequencing from Single Polymerase Molecules, Science, Vol.323, No.5910, pp.133-138 (2009) .
8) Dolezel, J., et al. : Nuclear DNA Content and Genome Size of Trout and Human, Cytometry A, Vol.51, No.2, pp.127-128 (2003) : author reply 129.
9) Wheeler, D.A., et al. : The Complete Genome of an Individual by Massively Parallel DNA Sequencing, Nature, Vol.452, No.7189, pp.872-876 (2008) .
10) Wang, J., et al. : The Diploid Genome Sequence of an Asian Individual, Nature, Vol.456, No.7218, pp.60-65 (2008) .
11) Ahn, S.M., et al. : The First Korean Genome Sequence and Analysis: Full Genome Sequencing for a Socio-ethnic Group, Genome Res, Vol.19, No.9, pp.1622-1629 (2009) .
12) International HapMap Consortium : The International HapMap Project, Nature, Vol.426, No.6968, pp.789-796 (2003) .
13) Sidransky, E., et al. : Multicenter Analysis of Glucocerebrosidase Mutations in Parkinson's Disease, N Engl J Med, Vol.361, No.17, pp.1651-1661 (2009) .
14) Druley, T.E., et al. : Quantification of Rare Allelic Variants from Pooled Genomic DNA, Nat Methods, Vol.6, No.4, pp.263-265 (2009) .
15) Johnson, S.M., et al. : Flexibility and Constraint in the Nucleosome Core Landscape of Caenorhabditis Elegans Chromatin, Genome Res, Vol.16, No.12, pp.1505-1516 (2006) .
16) Mavrich, T.N., et al. : Nucleosome Organization in the Drosophila Genome. Nature, pp. 358-362 Vol.453, No.7193 (2008) .
17) Sasaki, S., et al. : Chromatin-associated Periodicity in Genetic Variation Downstream of Transcriptional Start Sites, Science, Vol.323, No.5912, pp.401-404 (2009) .
18) Segal, E., et al. : A Genomic Code for Nucleosome Positioning, Nature, Vol.442, No.7104, pp.772-778 (2006) .
19) Lister, R., et al. : Human DNA Methylomes at Base Resolution Show Widespread Epigenomic Differences, Nature, Vol.462, No.7271, pp.315-322 (2009) .
20) Manber, U. and Myers, G. : Suffix Arrays , A New Method for On-Line String Searches, SODA, pp.319-327 (1990) .
21) Burrows, M. and Wheeler, D. : A Block-sorting Lossless Data Compression Algorithm, Digital SRC Research Report (1994) .
22) Kärkkäinen, J. and Sanders, P. : Simple Linear Work Suffix Array Construction, ICALP, pp.943-955 (2003) .
23) Ko, P. and Aluru, S. : Space Efficient Linear Time Construction of Suffix Arrays, CPM, pp.200-210 (2003) .
24) Nong, G., Zhang, S. and Chan, W.H. : Linear Suffix Array Construction by Almost Pure Induced-Sorting, DCC, Vol.202, pp.193-202 (2009) .
25) Li, H. and Durbin R. : Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform, Bioinformatics, Vol.25, No.14, pp.1754-1760 (2009) .
26) Li, H. and Durbin, R. : Fast and Accurate Long-read Alignment with Burrows-Wheeler Transform, Bioinformatics, Vol.26, No.5, pp.589-595 (2010) .

（平成 22 年 4 月 19 日受付）

森下真一（正会員）moris@cb.k.u-tokyo.ac.jp

1960 年生．1983 年東京大学理学部情報科学科卒業．博士（理学）．IBM，スタンフォード大学，東京大学医科学研究所，同理学部情報科学科を経て，2003 年より東京大学大学院新領域創成科学科情報生命科学専攻および理学部生物情報科学科教授．2009 年よりグローバル COE「ゲノム情報ビッグバンから読み解く生命圏」を推進中．