2. 今後の HPC が貢献しうる社会的課題

2.1 創薬・医療

(1) 社会的貢献 ―健康で長寿な社会を目指して―

日本はこれから急速に更なる高齢化社会を迎え、国民の健康の増進はきわめて重要な国家的課題となる。健康の増進に資する画期的創薬・医療技術の創出には、その基盤として人体等における生命現象の理解が不可欠であるが、生命現象はあまりに多くの要素が絡み合って複雑に関係している現象であり、まさに、今後の HPC での計算能力の飛躍的増大が有効に活用される分野であると言える。

生物の遺伝情報の単位である遺伝子、および遺伝情報全体を意味するゲノムは、生命の設計図であるとよく言われる。21 世紀に入ってすぐにヒトの全遺伝配列が決定されたが、それから 10 年近く経過し、遺伝子配列計測技術は飛躍的に進歩した。以前であれば数年かかった全ゲノム解読が、一人のゲノムについて数週間程度で可能になってきており、個人が自分自身のゲノム配列を知り、それに基づく医療（テーラーメイド医療）を受けることができる個人ゲノム時代が目前に迫っている。そのような超高速ゲノム解析を可能とする次世代 DNA シークエンサー（DNA を短く断片化し、並列に処理して高速に読み取る装置）では、膨大な観測データが日々算出されており、そこから意味のあるデータに処理するためには莫大な計算が必要となる。例えば、がんゲノムにおいては、膨大な遺伝情報の中から、がんの種類に応じたゲノム上の特徴を見出す必要がある。更には、個々の遺伝子配列のみならず、異なる因子が複合的に関わる疾患では、複数の遺伝子がどのように連携しつつ働いているかを解明する遺伝子ネットワーク解析も、HPC の重要な応用分野となっている。

ゲノム解析により疾患に関わるメカニズムが明らかになってくると、それを制御する薬をいかに開発するかという課題に直面する。通常、薬は、開発から市販まで、10 年以上の時間がかかるが、近年、更に対象の複雑化ゆえ、開発期間が長期化する傾向にあり、新薬開発のためのコストが上昇している。実際、一つの新薬開発には数百億円規模の投資が必要と言われており、技術的革新を引き起こすブレークスルーが望まれている。そのような状況下で、HPC を利用した計算創薬が期待を集めている。次世代の HPC で想定される膨大な計算資源を利用できるようになると、これまで物理化学分野で蓄積されてきた分子シミュレーションや量子化学計算などの予測信頼性の高い方法を創薬に応用することが可能になり、画期的な技術革新が期待される。更には、計算量が膨大になるためにこれまで考慮できなかった現実に近い環境、すなわち、細胞環境やウイルス全体を対象に含めた計算も、創薬に貢献するようになるであろう。

加えて、創薬のみならず、ナノテクノロジーと生命分野の境界領域に位置するナノバイオ分野において、生命分子と関連した新しいものづくりの技術革新への貢献が想定される（3.1.3 項を参照）。具体的には、タンパク質を用いた次世代デバイスやバイオセンサー、界面活性剤分子が会合したミセルや脂質膜を用いたドラッグデリバリシステム（患部に薬剤を届ける仕組み）や生体親和性の高いインプラント（人造の骨や歯）、細菌汚染の足場となるバイオフィル

ム（細菌などの微生物の増殖の足場）形成を阻害する洗浄表面などである。これらは分子間相互作用レベルでの制御がなされており、多くの生命分子と関連した次世代ものづくりへの展開が期待される。

医療応用の分野では、分子レベルのミクロなスケールから細胞・臓器・脳・全身スケールに至るまで、幅広い時空間スケールを統合的に扱うマルチスケールシミュレーションが重要な役割を果たす。例えば、心筋梗塞や脳梗塞などにおいては、血流中における血栓形成の理解が重要であり、HPC を用いた血栓形成のシミュレーション法の確立が医療応用上の重要課題として積極的に進められている。血栓症のみならず、糖尿病など、血液内のさまざまなイベントを通じて引き起こされる疾患は数多く、その予測に対する社会的要請は大きい。患者個別の情報を用いた血液中での薬効評価が可能になれば、患者個人の状態に合わせた医療を創出することにつながる。

また、医療機器開発における HPC の応用も重要である。患者の Quality of Life（QOL：生活の質）を向上させる非侵襲治療法の開発は重要な課題であるが、その一つである超音波治療機器の開発においても、シミュレーションの果たす役割は大きい。低侵襲治療によって患者の QOL が向上すれば、患者の社会復帰が容易になり社会が活性化するだけなく、医療費の低減につながることが期待される。現在の創薬・医療において、計算機による開発支援は非常に重要な課題である。

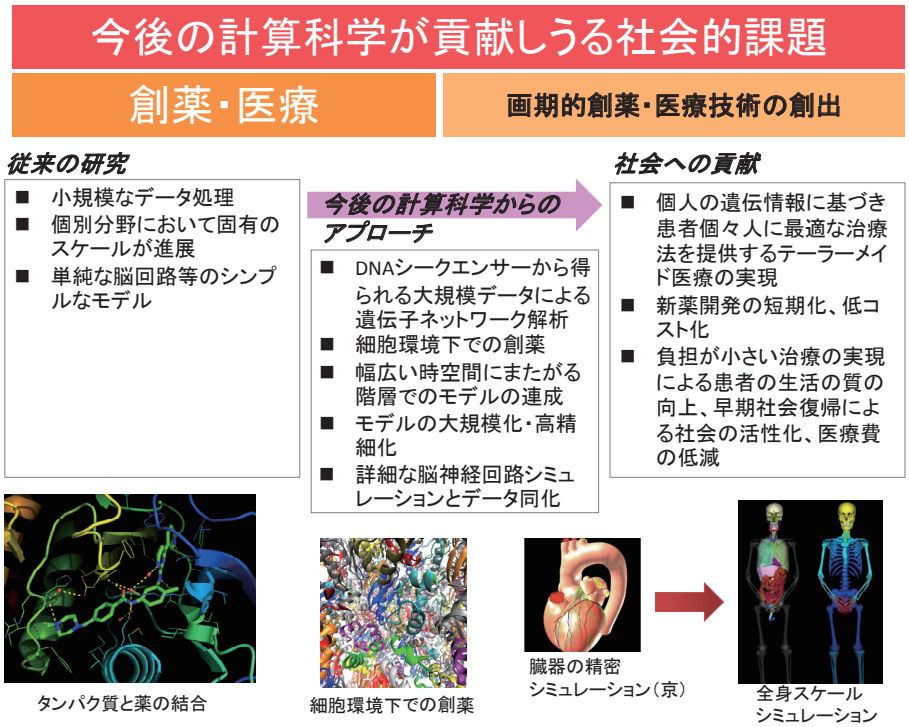

図 2.1-1　創薬・医療

(2)　サイエンスの質的変化

画期的創薬・医療技術の創出には、その基盤として人体等における生命現象の理解が不可欠であり、生命分野のサイエンスを強力に推し進める必要がある。生命分野は、その複雑さ故に

経験的色彩の濃い分野であったが、HPC の進展により、経験的方法論から、モデル化、シミュレーションといった演繹的方法論への転換が図られようとしており、まさに時代の転換点にさしかかっていると言える。生命現象のモデル化においては、原子・分子レベルのミクロなスケールから、細胞・臓器・脳・全身スケールに至るまで、時空間スケールが非常に幅広く、しかも各スケールが密接に関係し合っているという特徴がある。そこで、HPC の発展とともに、幅広いスケールを統合的に扱おうとするマルチスケール・マルチフィジックスシミュレーション法（幅広い時空間にまたがる対象に対し、それぞれの階層での計算を連成させるシミュレーション法）の開発が進展し、新たな分野を切り開きつつある。マルチスケール・マルチフィジックスシミュレーションでは、従来、別な分野で独立に進められていたさまざまなシミュレーションを統合することが求められる。分子レベルでは、ナノ分野で展開されてきた第一原理計算等の高精度の方法が生命分野に続々と応用され、大きな成果が期待される。また、細胞、細胞内小器官、ウイルス全体といった生命体の高次のレベルにまで、原子・分子のモデルを拡張したシミュレーション計算が可能になってくると思われる。また、そのような原子・分子レベルから、細胞レベル、臓器・全身スケールに至るまで、階層間をなめらかにつないでいくマルチスケール・マルチフィジックスシミュレーション技術を発展させることで、生命体の全体像が見えてくるであろう。その中で、特に発展が期待されるのが、脳・神経系である。人間の脳が、いかにしてこれほどまでに高度な知的情報処理を実現しているのか、その機構を解明することは、生命分野での究極の課題の一つである。人間の脳には 10^{11} とも言われる莫大な数の神経細胞が存在し、それらが 10^{15} のシナプスで複雑に結合していると推定されている。京の時代には不可能である人間の脳のシミュレーションも、次世代のエクサスケールの計算規模では、比較的単純な積分発火モデルであれば可能になると予想される。更に、昆虫の脳においては、エクサスケール級計算機を用いると全神経細胞を用いたリアルタイムシミュレーションが可能であると予想され、生理実験と連携して、神経回路網の機能の細部まで解明が進むと期待される。

また、今後の生命分野の計算科学に大きな影響を与えるものとして、実験施設から産出されるビッグデータが挙げられる。上記の次世代シークエンサーで高速に読み取られたゲノム情報の解析は、そのようなビッグデータ解析の代表例であるが（3.2.3 項を参照）、その他にも、X 線自由電子レーザー（3.3.1 項を参照）による生体資料の散乱像など、実験設備の高度化によって、そこから出てくるデータの量も莫大になっている。それを効率的に利用し、そこから意味のある情報を見出す技術開発が急務である。

以上のように、今後の HPC のもたらす莫大な計算能力が、さまざまな面で生命分野の発展に大きく資するのは間違いなく、ひいては画期的創薬・医療技術開発の重要な科学基盤となり得る。

(3) コミュニティからの意見

計算機科学の技術を利用して生物学の問題を解くバイオインフォマティクス分野からは、この分野のアプリケーションが、従来の典型的な HPC のアプリケーションとは異なる特徴があるという指摘があった。従来の典型的な HPC のアプリケーションでは、浮動小数点演算を重視しており、計算負荷の高い部分を高度に最適化して、繰り返しそのルーチンを使って計算するという方式をとるが、この分野のアプリケーションでは、膨大なデータを扱う都合上、スク

リプト言語の多用、アレイジョブ（同一のプログラムを異なる入力・設定で大量かつ同時に実行する並列計算の実行方法）、I/O、実数演算のみならず整数演算も重視するため、従来の HPC アプリケーションとは異なる評価基準を持つべきだと言う意見があった。また、現在のゲノム解析（3.2.3 項参照）を支えているヒトゲノム解析センターの現在および将来の計算能力が将来予想されるデータ規模に対して危機的な事が指摘された。

創薬分野からの意見では、他の大規模実験研究施設との連携についての指摘があった。創薬分野や分子レベルの計算では、タンパク質など生体分子の立体構造が重要な役割を果たしているため、それを決定する放射光施設などの大型実験研究施設の連携を十分に考慮しつつ、進展させていくべきだという意見がでた。更に、今後発展が期待される、X 線自由電子レーザー（波の位相が揃ったレーザーの性質を持つ超高輝度の X 線を発生することのできる光源）の研究施設である SACLA（国家基幹技術として兵庫県播磨に建設された X 線自由電子レーザー装置）（3.3.1 項を参照）とも十分によく連携していくべきという意見が出た。更には、今後のスーパーコンピュータで想定される、膨大な CPU コア数や計算を加速するアクセラレータを活用する計算方法やソフトウェアの開発についての意見も出た。

医療分野からは、医学系の分野では、HPC になじみがある人はまだ少数であり、今後とも HPC 分野と医学系分野の連携を深めていき、分野間のコミュニケーションを充実させるべきであるという意見が出された。

また、次世代研究者育成について、バイオインフォマティクス人材の不足、HPC 用プログラミング方法などの教育プログラムの充実や、生命分野で HPC の技術を持つ研究者の雇用先の拡充についての意見もあった。

（2012 年 10 月 19 日バイオスーパーコンピューティング研究会、2013 年 3 月 10 日文部科学省科学研究費補助金新学術領域研究「システム的統合に基づくがんの先端的診断、治療、予防法の開発」プロジェクトの公開講演会、2013 年 6 月 12 日～14 日　第 13 回日本蛋白質科学会年会、2013 年 6 月 27 日情報計算化学生物（CBI）学会研究講演会等にて）

(4)　計算機要求

上記のように、本分野は構成要素が非常に多いため、条件など詳細な個別の計算機需要は関連した箇所（3.1.3 項、3.2.3 項、3.3.1 項、4.1 節、4.2 節）を参照していただくとして、ここでは代表的なもののみを述べる。

まず、個人ゲノム解読を行う次世代シークエンサー解析については、2020 年頃に行われると思われる 200,000 人規模の解析を想定した。このようなビックデータ解析では、CPU 速度だけでなく、シークエンサーとストレージ間のネットワーク速度、メモリ容量、ストレージ容量・速度などデータに関連する性能の充実が望まれる。2020 年頃にはシークエンサーとストレージ間のネットワークの総和は 600～1200 GB/s、1 人のゲノム解析に要するメモリ容量は 1.6PB、演算性能は 5.4TIPS（tera-instruction per second）、ストレージ性能は 5.5TB/s、ストレージ容量は 100TB 程度必要となる。遺伝子ネットワーク解析については、2020 年には現在の 100 倍程度の量のデータが公開データベースにて入手可能になると予測され、そのデータ量を 4 万転写物×26,000 データセット・280 万アレイと想定して、ベイジアンネットワークおよび L1 正則化法の解析を行うとすると、恒常的に 25PFLOPS の計算需要が想定され、1 データセットの計

算に必要なネットワーク性能は 1.1PB/s, メモリ量は 0.08PB, ストレージ容量は 16TB 程度である。

創薬の分野でよく用いられる分子動力学（MD）シミュレーション（原子間相互作用に基づいて分子運動を計算する方法）では、短い時間間隔で繰り返し原子間相互作用を計算する必要があるため、演算性能重視でかつ高いネットワーク性能（特に低レイテンシ）が求められる。現在、標的タンパク質に対して 1000 種類の化合物の結合強度を調べると、10 PFLOPS の計算機を使えば 5 日くらいで計算が終了するが、1 EFLOPS の計算パワーがあれば、同じ時間内に、候補化合物をある程度網羅できる 10 万種類の化合物について標的タンパク質との結合強度が評価できる。ここでは、各々の化合物に対する結合強度は密に通信する並列ジョブで計算されるが、その並列ジョブは各化合物別のアレイジョブとして実行されるとした。また、複数のタンパク質を含む多成分系の超大規模計算へ拡張し、1 EFLOPS の実効演算能力があれば、細胞環境やウイルス全体や、細胞膜を模した脂質膜環境やドラッグデリバリシステムを全原子レベルで扱うことが可能な 1 億原子系の 1 μ 秒の MD シミュレーションを 2 日程度で完了可能である（以上の MD シミュレーションの詳細な見積もりは、4.1 節、4.2 節を参照いただきたい）。これにより膜輸送や細胞認識といった生命科学的現象を追跡することが可能となる。より長い分から時間スケールでの反応ネットワーク動態を再現する 1 分子粒度計算では、1EFLOPS の能力で 1,000 から 10,000 細胞で構成される細胞集団、あるいは組織における応答不均一性の計算が行える。更に、結合能計算の予測信頼性を上げるため、薬品分子（リガンド）と標的タンパク質との相互作用解析を量子化学計算で行った場合でも、1 EFLOPS の計算資源では候補化合物を絞るための 100 サンプルの同時処理が可能であり、計算効率が大きく向上すると期待できる。ここでは、水和条件下で 500 残基程度のタンパク質に一つのリガンドを組み合わせるケースを考え、10 PFLOPS の計算資源で 1 サンプル当たり 1 時間程度とした。また、10 PFLOPS ～1 EFLOPS の計算資源ではバイオデバイスで用いられる光関係タンパク質など、200～500 残基程度のタンパク質（電子軌道数は 10 万超）の実験分光データを日常的に解析するツールとなると思われる。以上の量子化学計算（FMO）で求められる計算スペックの詳細は、4.2 節を参照いただきたい。

医療応用の血流シミュレーションでは、詳しくは 4.1 節を参照いただきたいが、長さ 100mm、径 100 μ m の血管中を流れる血球の変形挙動を取り扱うために 0.1 μ m の格子幅を用い、血栓生成の時間スケールである 10 秒の現象を計算するには、必要な主記憶容量は 1 PB 程度である。全演算量は 2.5×10^{23}FLOP 程度となり、実効性能が 40%とした場合、1EFLOPS の計算機を用いると、約 174 時間の計算時間と見積もられる。ノード当たりの性能が 100 TFLOPS だとすると、理想的なキャッシュであれば B/F は 0.064 となるので、必要なメモリバンド幅は 6.4 TB/s と見積もられる。大規模な計算でモデルを精緻化することと併せ、10PFLOPS 程度の計算を 100 ケース程度並列させ、多くのパラメータスタディを行うことで医療に貢献するという視点もある。また、超音波シミュレーションでは、400 mm の立方体領域中の超音波の伝播が京の全ノードを占有した場合、実効性能が 20%程度なため、1 日程度で計算できる。しかし、軟らかい生体組織の温度変化で生じる微妙な剛性の変化をとらえるためには、より高い解像度と長時間の時間積分が必要になり、最低でも 1000 倍の自由度の計算となり、1 EFLOPS の計算資源で 10 日程の計算時間が必要である。また、マイクロカプセルとの干渉によって生じる超音波音場の解

析を 1 日で行うためには、計算コストが通常の超音波伝搬計算に比べて数十倍程度必要となるため、京の全ノードを占有した計算の 100 倍程度の計算処理速度が将来的に必要となると考えられる（詳しくは 4.1 節を参照いただきたい）。

脳神経系シミュレーションにおいては、2020 年ごろに実現可能なモデルとして一つには、積分発火モデルを用いた人の全脳規模の神経回路シミュレーションが考えられる。そのシミュレーションでは、1 秒のシミュレーションに 0.7 EFLOP 程度の計算量と 56PB のメモリ容量が必要であり、B/F 値はメモリに対しては 1 程度、ネットワークに対しては 0.25 程度が望ましく、700EFLOP 程度の計算量が見込まれる。

また、昆虫脳に対しては詳細なマルチコンパートメントシミュレーションが全脳に対して可能になり、更にシミュレーションを用いたパラメータ推定や実験とのリアルタイム通信による回路推定や制御が行われると見込まれる。カイコガの全脳における神経回路のリアルタイムシミュレーションの計算では、10%の実効効率を見込んで 700EFLOPS の計算性能と 0.83 のメモリ B/F が必要となる。全体の計算量としてはシミュレーションの開発と実行に 20EFLOP、更にその 100 倍程度の計算量で神経回路のパラメータ推定を行えることが見込まれる。更に、リアルタイムシミュレーションにおいては、シミュレータと生理実験装置と 100 MB/s 程度の通信が可能である必要がある。

課題	要求性能 (PFLOPS)	要求メモリバンド幅 (PB/s)	メモリ量/ケース (PB)	ストレージ量/ケース (PB)	計算時間/ケース (hour)	ケース数	総演算量 (EFLOP)	概要と計算手法	問題規模	備考
個人ゲノム解析	0.0054	0.0001	1.6	0.1	0.7	200000	2700	シーケンスマッチング	がんゲノム解析200,000人分のマッピングおよび変異同定	1人分の解析を1ケースとした。入力データを分割することで、細かい単位での実行、拠点をまたいだ実行も可能。整数演算中心のため「総演算量」はInstruction数とした。総浮動小数点演算量は45.864EFLOPとなる。
遺伝子ネットワーク解析	25.0	89.0	0.08	0.016	0.34	26000	780000	ベイジアンネットワークおよびL1正則化法	4万転写物×26,000データセット・280万アレイ	
創薬などMD・自由エネルギー計算	1000	400	0.0001	0	0.0012	1000000	4300000	全原子分子動力学シミュレーション	ケース数：10万化合物x10標的蛋白質(10万原子程度)	B/F=0.4, 数百から数千ケース同時に実行することを想定しているので，実行時に必要な全メモリ量，各ケースの実際の実計算時間は，表の値の数百～数千倍となる。メモリ量/ケースは100ノード実行時を想定。
細胞環境・ウィルス	490	49	0.2	1.2	48	10	850000	全原子/粗視化分子動力学シミュレーション	～1億粒子	B/F=0.1
細胞内信号伝達経路シミュレーション	42	100	10	10	240	100	3600000	一分子粒度細胞シミュレーション（格子法）	1000 から 10,000 細胞K構成される細胞集団	格子法:整数系の演算性能を要求。ケース数は最低10回、100回程度が望ましいため100回とした。
高精度創薬	0.83	0.1	1	0.001	1	100	300	薬品とタンパク質間相互作用の量子化学計算	水和条件下，500残基タンパク質＋リガンド	ファイルI/Oは終了時に1TBを1秒で書き出すことを想定し、1TB/s必要とした
バイオデバイス設計	1.1	0.2	1	0.001	1	100	400	200–500残基程度のタンパク質の分光計算	電子軌道数10万超	ファイルI/Oは終了時に1TBを1秒で書き出すことを想定し、1TB/s必要とした
血流シミュレーション	400	64	1	1	170	10	2500000	差分法、準陽解法(構造・流体・生化学連成シミュレーション)	100mm長x100um径，0.1um格子，流速10^−2m/s，解像度1us，10秒	
超音波シミュレーション	3800	4600	540	640	240	10	33000000	差分法、陽解法(音波・熱シミュレーション)	400mm^3の計算領域を軟組織とマイクロカプセル干渉音場を捉えるため、2250兆点の格子と時間ステップ数として1459200ステップが必要である。また、1格子点あたり演算数1000程度となる。	
脳神経系シミュレーション(ヒト全脳簡約モデル)	※7	※7.6	※56	※3600	0.28	100	700	単一コンパートメントIFモデル シナプス可塑性・通信	1000億ニューロン ニューロンあたり1万シナプス 10^5step	ネットワークのボトルネックはレイテンシー
脳神経系シミュレーション・昆虫全脳詳細モデル 神経回路パラメータ推定・生理実験とシミュレーションの通信	※71	※60	※0.2	※20	28	20	140000	マルチコンパートメントH-H(局所クランクニコルソン) シナプス通信 進化的アルゴリズム	1000ニューロン 10^6遺伝子 100世代	1OOMB/S程度の外部との通信も想定

※印の値は未だ精査中である。より精度の高い数値はWeb版（→「1.2. 本文書の構成」）を参照のこと。