ライフ分野のHPCと今後

次世代生命体統合シミュレーション研究開発プロジェクト
ISLiM: Integrated Simulation of Living Matter
姫野龍太郎
理化学研究所

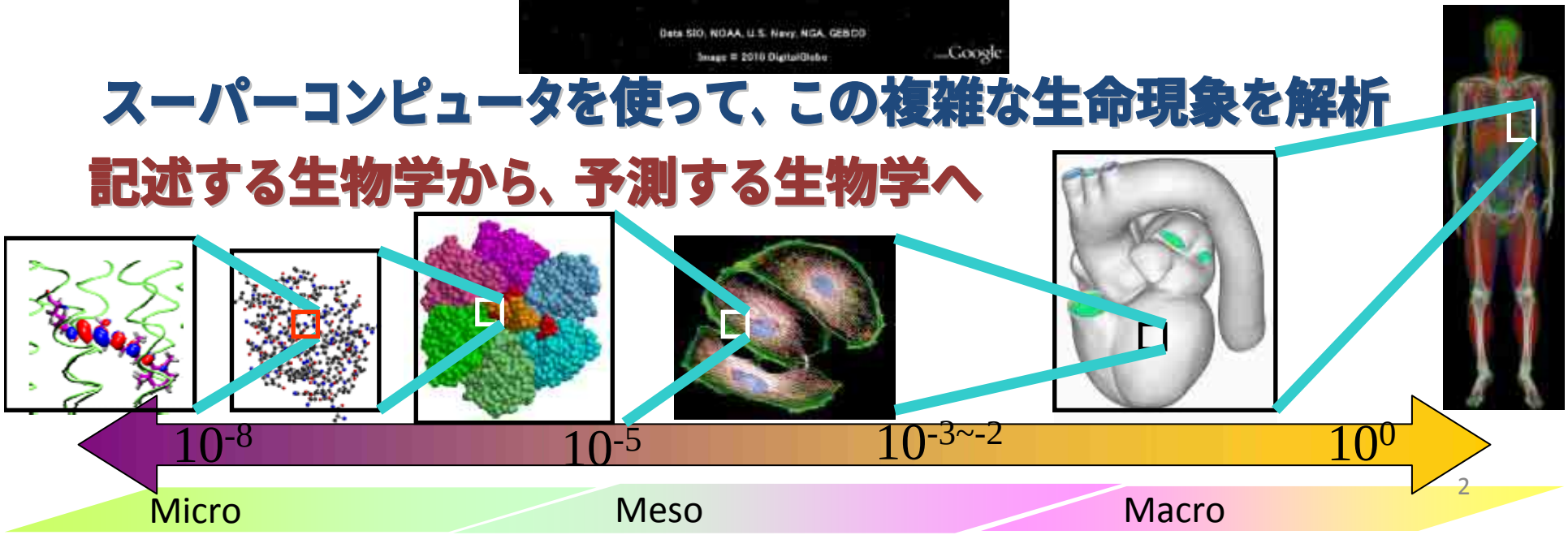
動機

生命現象は最も複雑で難しい問題

複雑で美しい振舞いを示す超多体系多階層問題

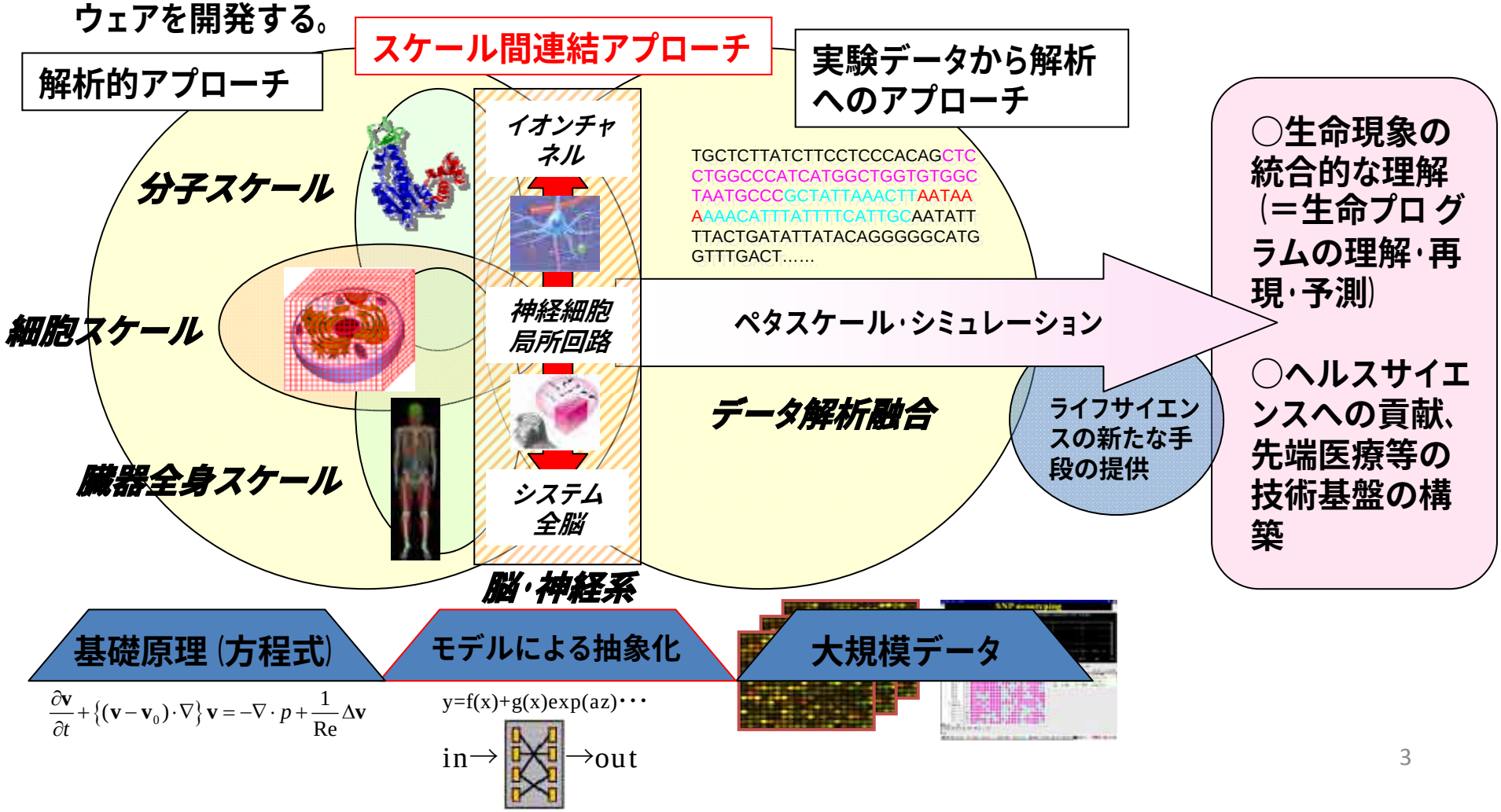
スーパーコンピュータを使って、この複雑な生命現象を解析

記述する生物学から、予測する生物学へ



研究開発の概要と達成目標

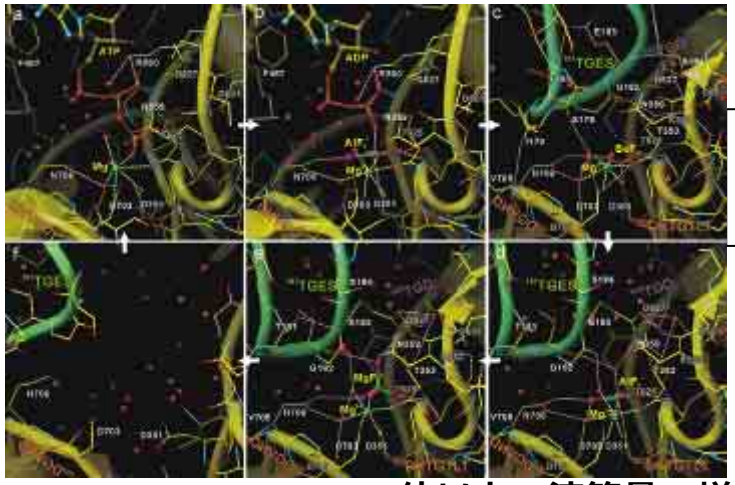
基礎方程式に基づく解析的アプローチと、大量の実験データから未知の法則に迫る実験データから解析へのアプローチ、さらには多階層を連結するアプローチにより、異なるスケールの研究と実験データを統合的かつ有機的に結びつけ、ペタスケールという桁違いの性能を持つスーパーコンピュータの性能をフルに発揮し、生体で起こる種々の現象を理解し医療に貢献するためのソフトウェアを開発する。



1) 分子シミュレーション

100T：全電子(QM)あるいはQM/MM計算に基づく蛋白質構造変化の解析（ナノ秒以上）

例）蛋白質内での酵素反応サイクルの第一原理動力学計算



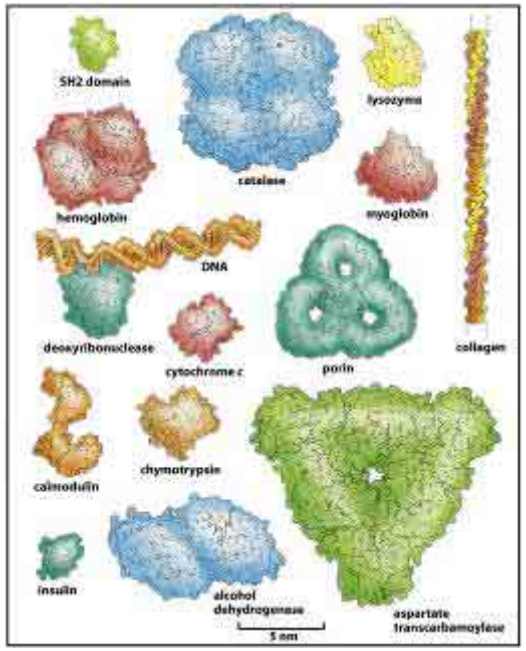
分子力場の改善
QM/MM法による長時間動力学
自由エネルギー評価法
1000倍の演算量の増加：100倍の性能向上＋計算方法の改良

10Peta：生体超分子複合体に関するミリ秒以上の分子動力学計算

例）生体超分子複合体の長時間分子シミュレーションの実現

モデルの粗視化
系の大規模化
計算の長時間（秒から分）

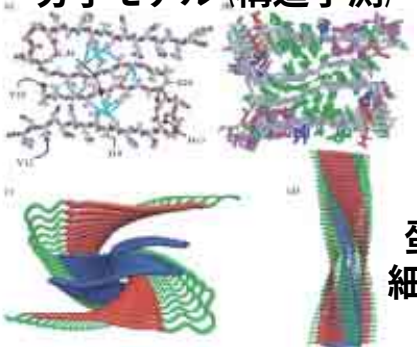
1000倍以上の演算量の増加を100倍のスパコン性能向上と計算アルゴリズムの改良で実現



1Exa：分子シミュレーションから細胞の動態解析へ（秒から分）

例）アルツハイマー病の原因と考えられているアミロイド凝集機構の解明

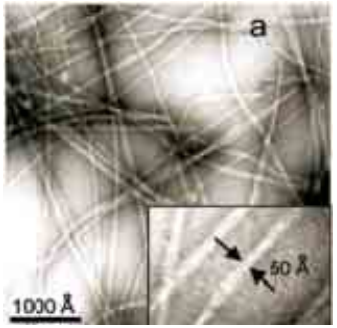
分子モデル（構造予測）



大規模計算
モデル粗視化

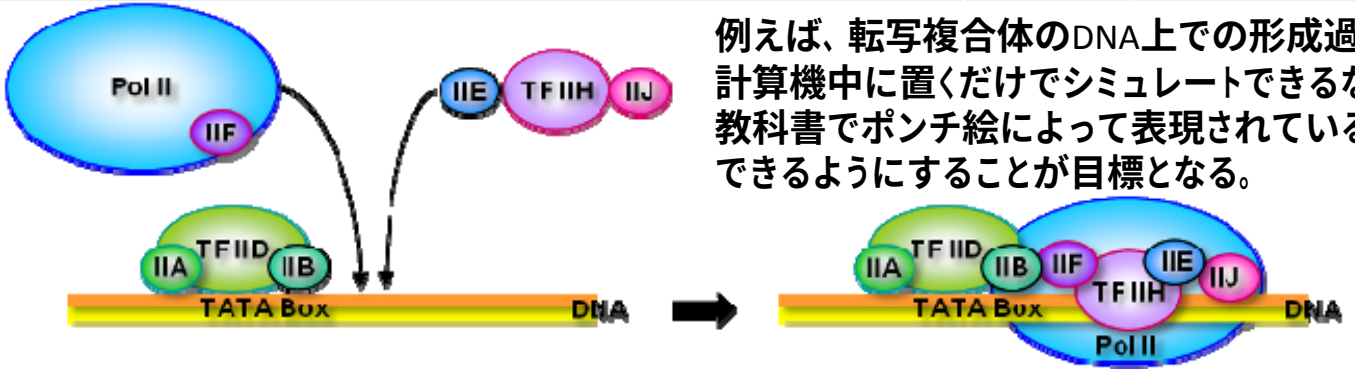
蛋白質の変化が細胞に及ぼす影響

細胞外でのアミロイド繊維の蓄積



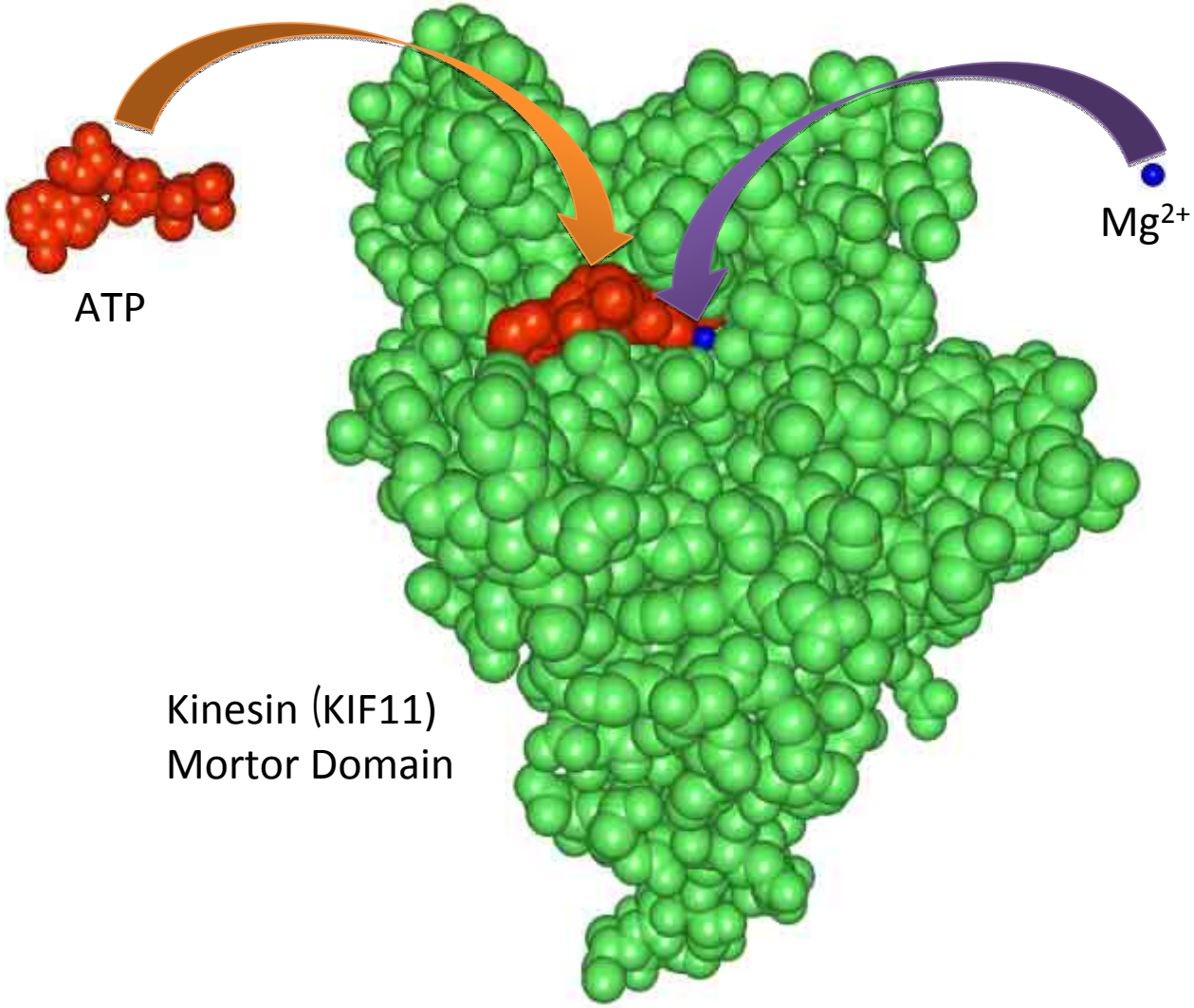
2) 分子シミュレーション

シミュレーション	＜ PetaFLOPS	10Peta FLOPS	＞ ExaFLOPS
種類	実験検証シミュレーション 予定調和的	→	予測シミュレーション 発見的
必要な実験情報	結晶構造解析データー 初期構造は結晶構造	→	実験情報は基本的に不要 初期構造はモデル構造
要素	希薄水溶液、脂質二重膜中の単一の生体分子	→	細胞モデル中の多数の生体分子
目的	平衡シミュレーション 単体分子のゆらぎ	→ 移行過程	非平衡シミュレーション 多数分子の会合･反応過程
範囲	限定的領域 結晶構造の周辺 (古典計算) 反応遷移状態周辺 (量子計算)	→	網羅的全領域 生状態で取り得るすべて (古典計算) 反応過程全体 (量子計算)
精度	限定的 経験的固定力場 (古典計算) 狭い量子領域 (QM/MM) 小さい基底、基底状態 (量子計算)	→	高精度 適応型力場 (古典計算) 広い量子領域 (QM/MM) 大きな基底、励起状態 (量子計算)



例えば、転写複合体のDNA上での形成過程を、必要な生体分子を計算機中に置くだけでシミュレートできるなどというように、生物学の教科書でポンチ絵によって表現されている現象を実際にシミュレートできるようにすることが目標となる。

具体例：Kinesinにおける拡散から解離までの ATP加水分解過程シミュレーション



1Exaでの概算の経過時間
(計算時間)

拡散
↓ 数10μ秒 (数-10時間)
遭遇
↓ 数10μ秒 (数-10時間)
活性複合体 (構造変化)
↓ 数m秒 (1ヶ月?)
加水分解 (QM/MM)
↓ 数μ秒 (1時間)
解離

~7-8万原子 (古典系)　　数千原子 (量子系)

分子スケールで課題となるHPC技術

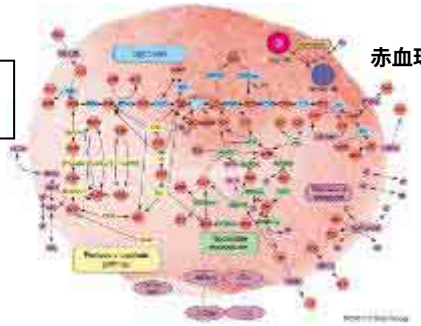
- 課題は現象を長時間追うこと
  - 専用機の威力：ANTONの衝撃
    - MD-GRAPE、ANTONなどの専用機
    - 現在の超並列以外の高速化の可能性
- 量子化学反応計算の高並列化
  - 大きな密行列、各ノードでは全体の一部を保持
  - ノードでどの部分を担当させるかのマッピング
  - 分散メモリー対応の行列演算ライブラリーは存在するが、利用しにくい
  - 使い勝手の良い、分散メモリー対応の行列演算ライブラリーの開発
  - バーチャル・シェアード・メモリーをサポートするハードやシステムソフト

2) 細胞シミュレーション

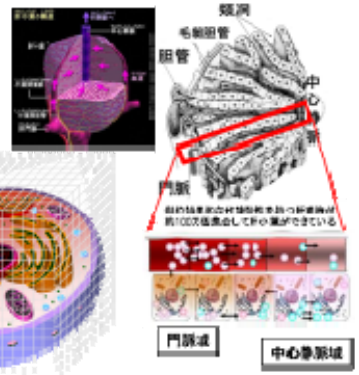
100T：細胞を1つの均一な空間と捉えたシミュレーション

例) 代謝シミュレーション
1細胞あたり約200酵素反応、400代謝物、数秒間の反応を計算
PC1台で数分間

10^6倍以上の演算量の増加：～10^6倍の性能向上＋並列化、効率化



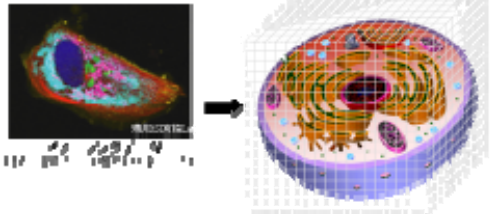
赤血球の代謝シミュレーション

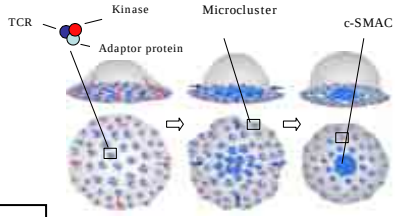


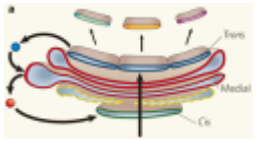
10Peta：細胞内の不均一な場を考慮したシミュレーション

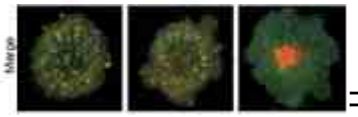
例) 細胞小集団 (肝小葉) の代謝シミュレーション
100x100x100ボクセル空間 (100nm分解能)
代謝反応、拡散反応、膜透過反応、物質輸送反応
1時間の反応を計算

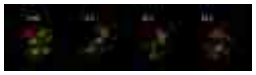
1000倍以上の演算量の増加：100倍の性能向上＋計算アルゴリズムの改良
生化学反応、構造･形態のダイナミクスに関する基礎方程式の確立











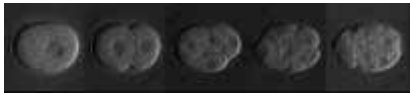
ゴルジ体の細胞内ダイナミクス

TCRミクロクラスタによる抗原認識と活性化の制御

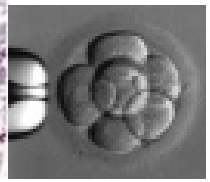
1Exa：細胞内の不均一な場の中での単分子のダイナミクス、細胞の構造･形態のダイナミクスを考慮したシミュレーション

例) 生化学統合シミュレーション (代謝反応、シグナル伝達、転写制御) 、胚発生･形成のシミュレーション
細胞内反応を基とした組織･臓器･器官の生化学血流連成シミュレーション

生化学ネットワークシミュレーションと構造･形態制御のシミュレーションの統合



初期胚の形態形成



代謝病、ガン、免疫疾患、再生医療などへの応用

例) ガン化のシミュレーション、組織再生のシミュレーション
シミュレーションを利用した予測、制御、設計

細胞スケールで課題となるHPC技術

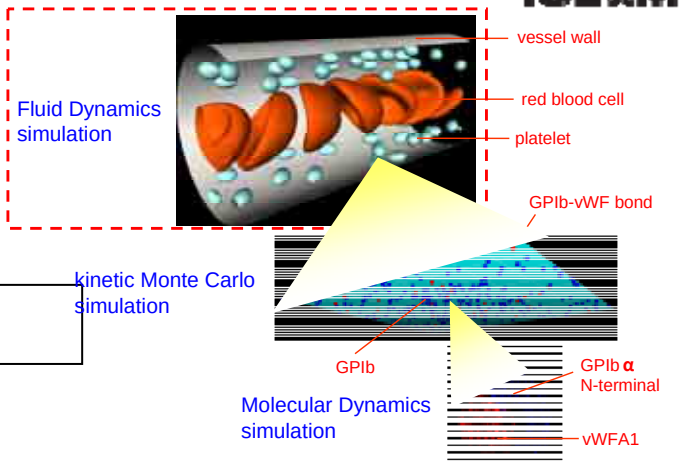
- 細胞スケールでのHPCは始まったばかり
- 現状での一番の課題は実験との比較検証と今後の応用における発展

- 並列化にミドルウェア：SPEHREを利用
  - 一個のノードでの計算と必要な隣接データを記述すると、自動的にMPI通信を含むコードを生成
  - 発展途上のソフトウェア開発には極めて有効

3) 臓器･全身スケール

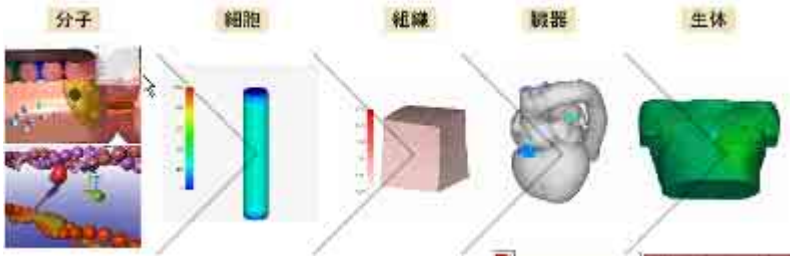
100Tera：単純化したモデルによる小スケールのシミュレーション

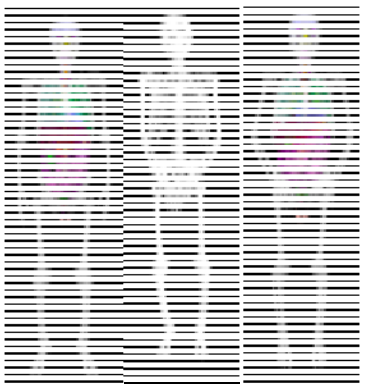
例1) 微小血管における1次血栓成長プロセスの初期課程
直径10ミクロンの血管内を赤血球と直径2ミクロンの血小板が流れ，血小板が血管壁に吸着するプロセスを再現.
例2) 低解像度均質化法を用いた心臓のシミュレーション
粗視化心臓による定性的再現，心疾患のモデリング
例3) 治療用超音波の焦点制御に関する低解像度計算



10Peta：細胞レベルからの臓器のフルシミュレーション

例1) 冠動脈に対する血栓成長･血管閉塞プロセスの再現
血小板表面の糖蛋白分子から，血球細胞，直径1mmの冠動脈内の血流まで考慮した血栓シミュレータ （10^12個の格子点の計算）
例2) 心筋細胞から心臓全体のフルシミュレーション
心筋細胞からの心臓のフルシミュレーション，
各種心疾患の再現，薬効の定量的評価，治療法の検討
(マクロ64万要素 X ミクロ20万自由度)
例3) 臓器全体を囲む計算領域での超音波腫瘍焼灼計算
実際の治療器の設計に利用可能な高解像度シミュレーション





1Exa：臓器と血流･神経系など複数の器官･組織の相互作用下での疾患の再現と薬効の評価･治療法の検討

例1) 詳細な生理学的因子を考慮にいれた動脈硬化から血栓形成･心筋梗塞にいたる細胞レベルからの心臓のフルシミュレーション
例2) タンパク質の変性による細胞スケールでの物性変化を考慮にいれた超音波治療の全プロセスシミュレーション

階層統合全身モデルによる生体の恒常性維持機構の解明と病態予測，薬効評価と治療法検討

臓器全身スケールで課題となるHPC技術

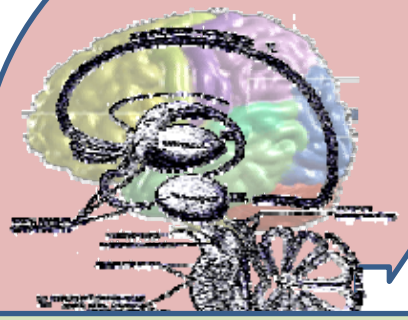
- ボクセルサイズ(要素サイズ)
  - 平均して各方向4倍、これが3次元で64倍、さらに時間方向のdtが1／4になるのでトータルで256倍の演算量、
  - 計算機は数百倍速くなるとすると、現実的な時間で計算可能
  - ボクセル数(要素数) は64倍なので、演算速度とメモリーの比はメモリーが少なくなって良い。
- 陰的な解法よりも陽的な解法の方が有利
- 流体部分はラティスボルツマンやシンプルなボクセル法にシフトして行く傾向にある
- 今後の傾向として
  1) 実験検証や実際に合うようなモデル化の方が困難であり、単純にスーパーコンピューティング技術だけで解決できる問題は少ない
  2) このような開発では、いろいろな試行錯誤をする必要があり、非常に大きなスーパーコンピュータよりも、手近なコンピュータで何度も試行錯誤することが重要で、PCやPCクラスターと巨大なスーパーコンピュータが同じプログラムで動くようにすることが必要。このようなハードやコンピュータシステムの差を吸収するライブラリーやディレクティブなどが必要 (この点ではSPHEREは有効)
  3) 効率的な開発のためには、計算結果を見るためのインタラクティブな可視化システムが必須

4) 脳神経系

データ統合・情報表現・心の科学

ライフサイエンス分野を超えて、社会科学全般に貢献

分子から行動までの統合



超大規模実時間シミュレーション

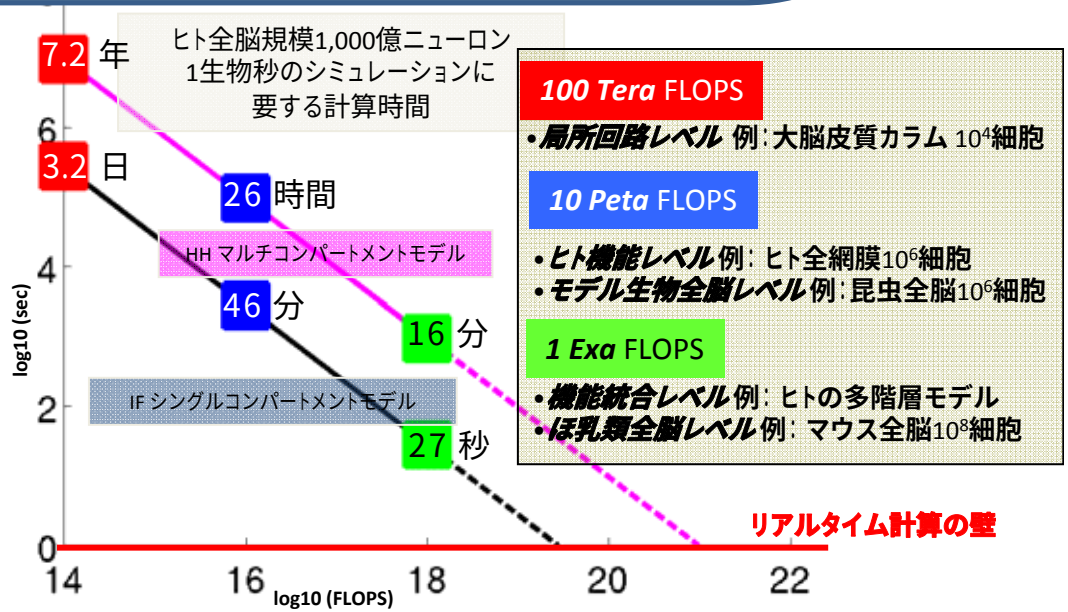
脳のネットワーク異常と考えられるうつ病や統合失調症などの神経疾患の理解

脳領域の統合

各階層のデータ（実験・シミュレーション）

神経細胞

分子シグナル伝達



脳神経系スケールで課題となるHPC技術

- 異なる脳領域（大脳皮質、大脳基底核、小脳、脳幹）を横に統合し、また分子から行動までを縦に統合した超大規模実時間シミュレーションが可能
- 脳のネットワーク異常と考えられるうつ病や統合失調症などの神経疾患の理解が進む。
- この際の、領野間の配線、あるいは局所回路内の詳細回路配線に関する情報は、拡散MRIや光計測に基づく機能的回路観測法などのコネクトミクス技術の進展により、現在データの集積が著しく、次次世代機の際には利用可能になっていると考えられる。
- データベースを取り込んだシミュレーション：高速分散並列IO

5)データ解析:ゲノムを基軸とした個別化医療への展開

100T：ゲノム革命の第一フェーズ

社会･テクノロジーの背景
パーソナルゲノム時代が幕開け（全世界で3万人のヒトゲノムデータ）。
日本では次世代シークエンサーが100台程度稼働。
医療情報の電子化はまだ夜明け前。高齢化社会への不安が高まる。

研究の状況
要素還元論的な生命システム･病気の理解の限界が認識される。
数百サンプルを用いた遺伝子と病気の統計的関連付けが日常的に行われる。
大規模データ解析による、がんや生命システムの複雑さの探索が開始される。

10 Peta：生体内分子や環境が織りなす時間軸・空間軸のある病態・生命システム、及びその多様性の理解

社会･テクノロジーの背景
次々世代、シリコンシークエンサーの登場。10万円でパーソナルゲノム。
パーソナルゲノム情報に対応した医療情報の電子化。
災害･救急に対応する情報システム。高齢化によるがん･生活習慣病の急増。
有効とされた薬剤が重篤な副作用で撤退。個別化医療が決定的。

研究の状況
個人のクリニカルシークエンスデータ解析･個人に合ったバイオマーカーの探索。
がんや病態の四次元的理解と病気のデジタル化。
1万円･1時間でパーソナルゲノムが解析可能。全人類規模で起こる大変革に、
災害にロバストな超大規模計算、大規模ストレージ、超高速ネットワークで対応。

1 Exa：個人個人のＤＮＡ情報と医療情報に基づいた個別化医療の実現

HPCインフラに支えられた、ゲノムが社会に融合した新社会が出現。
だれもが健康保険証を持っているようにパーソナルゲノムをもって医療を受ける時代。
個人のゲノム情報と医療情報が使える「スマートフォン」。予防医療、新たな医療サービス市場の登場。

ゲノムを基軸とした個別化医療への展開のために必要な計算量とデータ量 (①と②の実施)

		2012	2013	2014	2015	2016	2017	2018	2019	2020
ヒト生命データ解析	演算性能	250TFLOPS			1PFLOPS		256PFLOPS			1ExaFLOPS
	ストレージ	16PB			40PB		128PB	256PB	512PB	1ExB / 2ExB
	サンプル数	500	1,000	5,000	7,000	10,000	20,000	40,000	80,000	200,000〜1,000,000
		初期		中期			後期			
1ｻﾝﾌﾟﾙ当たりのﾃﾞｰﾀ		1TB		1TB〜10TB			10TB〜100TB			
シークエンサー技術		次世代シークエンサー		第三世代シークエンサー			第四世代シークエンサー			

① 個別化医療推進のための情報基盤の整備
- ヒト大規模ゲノム関連データベース (数十万人)、副作用情報データベース、大規模生命・医療データ解析技術、ソフトウェア等の情報基盤技術を整備。このために、スーパーコンピュータシステム (1エクサ・フロップス)、及び大規模ストレージ (2エクサ・バイト) がインフラ設備として必要となる。

② 個別化医療のための次世代ゲノム医学研究
- 超高速シークエンサー技術等を駆使して、個人個人のゲノム情報・エピゲノム・トランスクリプトーム・プロテオーム・メタボロームなどの違いと、がんや成人病等の病気、薬、環境因子との繋がりを解明し、それを診断、予防、治療へと翻訳する最先端研究の実施。

データ解析で課題となるHPC技術

- Exa Byte級の高速並列入出力システム
- IOがボトルネックにならないための全体システムの設計
- データベースを常駐化したシステム運用体制
- 人データを取り扱うためのセキュリティーの確保と研究倫理のあり方 (コンセンサス)
- データベースの保存、二重化、デザスタ対応

まとめ

- 生命科学分野では解析の対象によってＨＰＣでの課題は異なる
- 一つの解きにくい方程式を解くタイプの難しさではなく、スケールをまたぐ階層的な現象、未知な現象が潜む対象を取り扱う難しさ
- 複雑に絡み合う現象を統合的に取り扱う唯一の手段がＨＰＣ

- ソフトウェアの開発しやすさ、組み合わせの容易さ、ＰＣからスパコンまでが同じソフトで動く仕組み
- 大量のデータベース利用、人情報のためのセキュリティ管理