研 究 解 説

次世代量子化学計算システム

Quantum Chenmial Calculation System: the Next-Generation

佐 藤 文 俊*

Fumitoshi SATO

1. はじめに

タンパク質は生体の主要成分で，H, C, N, O, Sといった原子からなる20種のアミノ酸がペプチド結合により連結したポリペプチド鎖からなる，数百～数十万の原子を持つ巨大分子である．個々のタンパク質はアミノ酸配列順序，いわゆる一次構造が異なるだけでなく，高次構造も異なる．すなわち，α－ヘリックス，β－構造，ランダムコイルといった局所的な二次構造や，これらが空間的に折りたたまれて形成される三次構造，さらに複数のポリペプチド鎖からなるタンパク質では特有の四次構造を持つ．そのような高次構造の形成により，配列上離れているアミノ酸を三次元的に接近させて，機能領域である活性部位を形成する．

また，全てのタンパク質のうち約1/3は何らかの形でNa, K, Mg, Ca, Mn, Fe, Co, Ni, Cu, Zn, Se, Moといった金属イオンを必要とし，その大部分は生体に必須な金属である．これら金属イオンはタンパク質構造の安定性に寄与するだけでなく，ヘム（鉄ポルフィリン）・クロロフィルのような補欠分子族や，鉄－硫黄クラスタのような原子集団として組み込まれ，活性中心そのものを形成する．これらは，ヘムタンパク質（口絵図1），光合成反応中心，鉄－硫黄タンパク質に代表されるように，非常に活性が高く，また様々な反応性を示す．ヘムタンパク質に限っても，酸化還元，電子伝達，酸素運搬，酸素添加などその機能は多岐にわたっており，ヘムの様々なスピン状態をタンパク質部分が制御していることを示している．

その上，タンパク質は生体内で働くため，体温で反応が進行する必要がある．このことは，タンパク質がわずか0.1 eV程度のエネルギー変化で機能する分子であることを意味する．

このように，タンパク質の機構の研究には，タンパク質部分の研究に加えて金属イオンを含めた活性中心の研究が重要であり，これらが巨大分子系という困難の下で，精度よく解析されなければならない．実験的研究においては，様々な線源による分光法などの解析装置の進歩やその解像度の格段の向上，タンパク質精製技術の発展，そして遺伝子操作による機能解析法の登場により，分子生物学として大成功を収めている．

一方，理論的研究では，これまで分子のサイズがネックとなり，古典論やモデルによる解析方法が主であった．しかし，タンパク質が関与する様々な化学反応は，有機化学や無機化学でもそうであるように，電子状態が重要な役割を担っている．有機・無機化学の分野で大成功を収めた理論は量子化学であり，その手法は分子軌道法と呼ばれている．この信頼できる手法をタンパク質に適用し，タンパク質そのものの電子状態を明らかにすることは，まさに究極の理論分子生物学といえるであろう．

当グループでは，アミノ酸残基と金属イオンで同程度の定量性を持ち，かつ数千から数万原子からなるタンパク質をありのまま扱う量子化学計算による解析法の開発が欠かせないと考えた．定量的な分子軌道計算のためには電子相関効果が重要である．巨大系をまるのまま扱うには計算量がポイントとなる．現在の計算機事情では，これらの要件を満たす手段として，分子軌道法のスタンダードである*ab initio* Hartree-Fock（HF）法[1]と同程度の計算量で，電子相関を取り込むことができるKohn-Sham-Roothaan（KSR）方程式に基づく密度汎関数法[2]が現実的である．

そこで当グループでは，タンパク質のための密度汎関数法プログラムProteinDFを開発し[3]，これをワークステーションクラスタ上に用いてシトクロムc（口絵図1）の全電子計算に世界で始めて達成した[4,5]．現在，このProteinDFをベースに機能を大幅に追加し，インフラを整備して，次世代量子化学計算システムとして発展させている．タンパク質の電子状態計算の実用化に貢献するとともに，基礎研究のみならず産業界においても有用なツールに育て上げたいと努力している．本稿では，ProteinDFを中心に，次世代量子化学計算システムについて紹介する．

*東京大学生産技術研究所　計算科学技術連携研究センター

2. 密度汎関数法

2.1. KSR方程式

物質の電子状態は，基礎方程式であるSchrödinger方程式の解を求めることにより得られる．しかし，今日でも，Schrödinger方程式を多原子系で解くことは大変困難で，種々の近似が行われている．その一つが，多電子波動関数を1つのSlater行列式で表し，その1電子軌道を求めるというHF近似で，これが現在でも最も標準的な方法である．

しかし，この近似は多電子効果（電子間の相関効果）が考慮されていないため，特に金属を含んだ系では精度のよい解を与えない．多電子効果を取り入れるすなおな拡張は，Slater行列式の線形結合を導入する配置間相互作用法[1]であるが，計算量が系のサイズの5乗に依存し，大きな系に向いていない．

そこで，その解決方法の1つとして，Slater, Kohn, Hohenberg, Shamらの理論[6-8]により，単一のSlater行列式の立式において，電子相関の効果をポテンシャルに繰り込んだ演算子を導入する密度汎関数法が発展してきた．

密度汎関数法では全エネルギーを電子密度ρの汎関数で表す．最適なρの組を見つけるためにスピン軌道の規格直行性の束縛条件下で，全エネルギーをρについて変分をとると，Kohn-Sham（KS）方程式が得られる．ただし，ここでは簡単のためα，βスピンの分子軌道が同じ場合を示す．

$$\left[-\frac{1}{2}\Delta-\sum_A\frac{Z_A}{|\boldsymbol{r}-\boldsymbol{R}_A|}+\int\frac{\rho(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}d\boldsymbol{r}'+\mu(\boldsymbol{r})\right]\phi_i(\boldsymbol{r})=\varepsilon_i\phi_i(\boldsymbol{r}) \quad \cdot(1)$$

ここで，左辺括弧内はKS演算子と呼ばれ，左から運動エネルギー，電子核間引力，電子間反発，交換相関ポテンシャルの各演算子を表す．ε_iはKS固有値（軌道エネルギー）である．この方程式はHF方程式とよく似ていて，実際$\mu(\mathbf{r})$を交換演算子に置き換えれば，HF方程式と同型となる．

この方程式を数値的に解く方法もあるが，現在とのころ多原子分子では実用的ではない．そこで，分子軌道に既知の空間基底関数の組（g_p; p=1, 2,…, N_p）を導入して代数方程式の組に変換し，これを行列方程式として解く方法が用いられる．具体的には以下のように展開する．

$$\phi_i(\boldsymbol{r})=\sum_p C_{pi}\,g_p(\boldsymbol{r}) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots(2)$$

理想的にはg_pが完全系であれば正確な展開であり，関数も任意である．例えば，基底関数に平面波を用いるCar-Parrinello法[9]も提唱されている．残念ながら，実際の計算では展開は有限個で打ち切らざるを得ず，有限基底関数が張る部分空間においてのみ正確である．そのような意味で，十分に精度のよい展開を与える基底を選ぶことが必要といわれている．通常，量子化学では基底として原子軌道を模していて，かつ計算が楽なガウス型関数を組み合わせて使用する．そのため，C_{pi}はしばしばLCAO（Linear Combination of Atomic Orbitals）係数と呼ばれる．

式（2）と同様に電子密度，交換相関ポテンシャルにも（補助）基底関数展開を導入し，

$$\rho(\boldsymbol{r})\cong\tilde{\rho}(\boldsymbol{r})=\sum_\alpha\rho_\alpha\,g^\rho_\alpha(\boldsymbol{r})$$
$$\mu(\boldsymbol{r})=\sum_\gamma\mu_\gamma\,g^\mu_\gamma(\boldsymbol{r}) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots(3)$$

これらを式（1）に代入すると，KSR方程式が得られる．

$$\boldsymbol{FC}=\boldsymbol{SC}\boldsymbol{\varepsilon} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots(4)$$
$$F_{pq}=h_{pq}+\sum_\alpha\rho_\alpha\langle pq|\alpha\rangle+\sum_\gamma\mu_\gamma\langle pq\gamma\rangle$$
$$h_{pq}=\int g_p(\boldsymbol{r})\left[-\frac{1}{2}\Delta-\sum_A\frac{Z_A}{|\boldsymbol{r}-\boldsymbol{R}_A|}\right]g_q(\boldsymbol{r})\,d\boldsymbol{r}$$
$$\langle pq|\alpha\rangle=\iint g_p(\boldsymbol{r})g_q(\boldsymbol{r})\left[\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}\right]g^\rho_\alpha(\boldsymbol{r}')\,d\boldsymbol{r}d\boldsymbol{r}'$$
$$\langle pq\gamma\rangle=\int g_p(\boldsymbol{r})g_q(\boldsymbol{r})g^\mu_\gamma(\boldsymbol{r})\,d\boldsymbol{r}$$
$$S_{pq}=\langle pq\rangle=\int g_p(\boldsymbol{r})g_q(\boldsymbol{r})\,d\boldsymbol{r} \quad \cdots\cdots\cdots(5)$$

ここで，$\boldsymbol{F}, \boldsymbol{C}, \boldsymbol{S}, \boldsymbol{\varepsilon}$は行列で（$\boldsymbol{\varepsilon}$は対角行列），$\boldsymbol{C}, \boldsymbol{\varepsilon}$は$\boldsymbol{F}$を対角化して求まる固有ベクトルと固有値である．ただし，基底関数は規格化されているが，直交はしていないため，$\boldsymbol{S}$は単位行列ではない．そこで，$\boldsymbol{S}$を対角化するユニタリー行列を使って，直交変換行列$\boldsymbol{X}$を作る．

$$\boldsymbol{X}=\boldsymbol{U}\boldsymbol{s}^{-\frac{1}{2}}$$
$$\boldsymbol{U}^\dagger\boldsymbol{S}\boldsymbol{U}=\boldsymbol{s} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots(6)$$

ここで，$\boldsymbol{s}$は$\boldsymbol{S}$の固有値を対角要素に持つ対角行列である．-1/2乗は対角要素の平方根の逆数をとることを意味する．従って，

$$\boldsymbol{F}'\boldsymbol{C}'=\boldsymbol{C}'\boldsymbol{\varepsilon}$$
$$\boldsymbol{F}'=\boldsymbol{X}^\dagger\boldsymbol{F}\boldsymbol{X}$$
$$\boldsymbol{C}'=\boldsymbol{X}^\dagger\boldsymbol{C} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots(7)$$

これは普通の固有値問題である．

以上を，ガウス型基底関数を用いた密度汎関数法と呼ぶ．この方法は，$\boldsymbol{F}$の生成以外はHF方程式の解法（*ab initio*

HF法）とほぼ同じであるといってよい．このように，配置間相互作用法に比較して，密度汎関数法は計算量が *ab initio* HF法並と大幅に少ないにもかかわらず，電子相関を有効に取り入れて，水素や炭素などと金属原子をほぼ同等の精度で計算することができるのが最大の魅力で，タンパク質の定量的な計算に最も適した方法である．現在でも，密度汎関数法における交換相関ポテンシャル演算子をさらに改善していくことは最先端の研究の1つであり，これからもますます精度が向上するものと見込まれている．

2.2. 密度汎関数法の計算手順

式（1）はKS演算子がこの固有値方程式の解である ϕ_i に，電子間反発演算子と交換相関ポテンシャル演算子を通して依存しているので，非線型方程式である．従って，繰り返しの方法によって解かねばならない．これを自己無撞着（Self-Consistent Field; SCF）法と呼ぶ．そこで，適当な初期値 C_{pi} または ρ_α から計算を出発する．

C_{pi} から出発する場合，これから電子密度行列

$$P_{pq}=\sum_{i=1}^{N/2}2C_{pi}C_{qi}^{*} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots(8)$$

を計算する．次に，クーロンエネルギーの差が最小になるよう電子密度の展開係数をフィッティングする．

$$\rho_\alpha=\sum_\beta\langle\alpha|\beta\rangle^{-1}\left\{\sum_{pq}P_{pq}\langle pq|\beta\rangle-\Lambda\int g_\beta^{\rho}(\boldsymbol{r})d\boldsymbol{r}\right\}$$

$$\langle\alpha|\beta\rangle=\iint g_\alpha^{\rho}(\boldsymbol{r})\left[\frac{1}{|\boldsymbol{r}-\boldsymbol{r}'|}\right]g_\beta^{\rho}(\boldsymbol{r}')d\boldsymbol{r}d\boldsymbol{r}' \quad \cdots\cdots(10)$$

ここで，-1乗は逆行列を意味し，Λ はLagrangeの未定乗数である．続いて，交換相関ポテンシャルおよびエネルギー（ε^{XC}）の展開係数をフィッティングする．

$$\mu_\gamma=\sum_\delta\langle\gamma\delta\rangle^{-1}\int\mu(\boldsymbol{r})g_\delta^{\mu}(\boldsymbol{r})d\boldsymbol{r}$$

$$\varepsilon_\gamma^{XC}=\sum_\delta\langle\gamma\delta\rangle^{-1}\int\varepsilon^{XC}(\boldsymbol{r})g_\delta^{\mu}(\boldsymbol{r})d\boldsymbol{r}$$

$$\langle\gamma\delta\rangle=\int g_\gamma^{\mu}(\boldsymbol{r})g_\delta^{\mu}(\boldsymbol{r})d\boldsymbol{r} \quad \cdots\cdots\cdots\cdots\cdots(11)$$

これで式（5）に従って，F_{pq} を作ることができる．後は式（7）により規格直交基底に変換し，固定値問題を解き，新しい C_{pi} を得る．この操作を，SCF計算の前後で対応する行列全要素の差の絶対値や全エネルギー

$$E=\sum_{pq}P_{pq}\left\{h_{pq}+\sum_\alpha\rho_\alpha\langle pq|\alpha\rangle+\sum_\gamma\varepsilon_\gamma^{XC}\langle pq\gamma\rangle\right\}$$

$$-\frac{1}{2}\sum_{\alpha\beta}\rho_\alpha\rho_\beta\langle\alpha|\beta\rangle+\frac{1}{2}\sum_{AB}\frac{Z_AZ_B}{|\boldsymbol{R}_A-\boldsymbol{R}_B|} \quad \cdots\cdots(12)$$

の差が閾値以下になるまで繰り返す．これを収束と呼ぶ．

以上をまとめると以下の通りである．

$$\ldots\rightarrow\tilde{\rho}\rightarrow\mu\rightarrow\boldsymbol{F}\rightarrow\boldsymbol{F}'\rightarrow\boldsymbol{C}',\varepsilon\rightarrow\boldsymbol{C}\rightarrow\boldsymbol{P}\rightarrow E\rightarrow\ldots \quad \cdots\cdots\cdots(13)$$

先に述べたように，ガウス型基底関数を用いた密度汎関数法は *ab initio* HF法とよく似ており，長い間培われてきたノウハウを利用することができる．

(1) 式（11）上2つ以外の積分（総称して分子積分と呼ぶ）は，ガウス関数の多中心積分がやはりガウス関数である事実を使って，全て解析的に求めることができる．式（11）上2つは被積分関数が有理関数であるため，数値積分で求めている．

(2) $\langle pq|\alpha\rangle$，$\langle pq\gamma\rangle$ は3中心積分のため，見かけ上は系のサイズ（N）の3乗に依存する計算量を有するが，ガウス型関数は距離が離れるとすぐに値が小さくなることを利用して，大きな分子では計算量を N の自乗程度に抑えることができる．

(3) 分子積分以外は，行列の積や逆行列，対角化といった一般行列演算により計算される．

(4) 通常，規格直交基底では計算に登場する行列はほとんど素行列にならないので，行列演算は N の3乗の計算量が必要である．大きな分子ではこれが問題となる．

(5) 式（13）のSCFを収束させるためには初期値が良いことが絶対条件だが，収束を援助または加速する様々な技術（ダンピング法，DIIS法，レベルシフト法，射影演算子法，アップデート法など）が開発されている．

3. オブジェクト指向技術とProteinDF

複雑なプログラムを様々な計算機環境に柔軟に対応するために，当グループのプログラム開発にはオブジェクト指向技術を導入した．これにより，多人数によるプロジェクト型コーディングがスムースに行われ，プログラムの並列化やベクトル化による高速化が容易となった．

オブジェクト指向プログラミングの基本的な構造は「オブジェクト」，「クラス」，「属性」，「メソッド」，「メッセージ」の5つの用語と，それに伴う「カプセル化」，「抽象化」，「継承」という3つの概念で説明される．

プログラムはオブジェクトが中心となって構成される．個々のオブジェクトはその作業である役割を果たすために必要な属性（データ）とメソッド（手続き）を内部に持っている．これをカプセル化と呼ぶ．この大きな特徴を利用して，オブジェクトの提供者と利用者を明確に切り分けることができる．例えば分子積分を行うオブジェクトを利用する人物は，分子積分オブジェクトの内部仕様を知る必要

がなく，またこれを作成する人物も自身が提供する呼び出し手続きの仕様さえ決めておけば，まったく独立に構築することができる．さらに，この呼び出し手続きの仕様を変更しなければ，オブジェクト内部の変更は問題なく行うことができる．つまり，利用者のプログラムを変更することなく，例えばワークステーションクラスタ用分子積分オブジェクトとベクトル計算機用分子積分オブジェクトの好きなほうを使用できる．これを仕様と実装の分離と呼ぶ．

オブジェクトどうしはメッセージを送りあうことで，協調的に作業を進める．クラスはオブジェクトのための鋳型であり，プログラマが実際にコーディングするプログラムで，これから多数の同型のオブジェクトを生み出すことができる．これは抽象化という概念と関係がある．例えば複数のマシンがメッセージをやり取りしながら並列に計算を進めていく環境を，マシンオブジェクトがメッセージをやり取りしてプログラムを実行する形に表現することができる（後述；図3参照）．これにより，様々な計算機環境に無理なく対応できるプログラム構造となる．

また，オブジェクト指向では，上位クラスの概念を下位クラスが引き継ぐとともに新たな属性やメソッドを追加できる．これを継承という．クラスを体系化する手段，既存の資源を有効に再利用して，生産性を向上させる手段である．当グループでは，これをファイルの共通基本操作を上位クラスに，これを継承しデータ構造に依存する特定の操作を追加して下位クラスを構造化した．また，既存の資産に傷つけることなく拡張できるので，複数人によるプログラム開発に適している．

図1にProteinDFプログラムの構造を示す．C++[10]を採用し，計算の単位をクラスにするモデルを採用した．

計算が開始されると入力解析（DfInputdata），コアの分子積分（DfIntegrals），初期値作成（DfInitialguess），そして，SCF計算（DfScf）が行われ，計算が終了する．

DfScfのオブジェクトは様々なオブジェクトを逐次生成させて計算処理を実行する．規格直交基底変換行列計算（DfXmatrix）後，交換相関ポテンシャルフィッティング（DfGrid），KS行列生成（DfFockmatrix），KS行列の規格直交基底変換（DfTransFmatrix），対角化（DfDiagonal），LCAO係数の逆変換（DfTransatob），密度行列計算（DfDmatrix）を経て全エネルギーを計算する（DfTotalenergy）．ここで，収束判定（DfConvcheck）を行い，収束したらDfScfが終了，収束条件を満たさない場合は電子密度フィッティング（DfDensityfitting）が行われ，DfGridに戻る．DfEriは$\langle pq \mid \alpha \rangle$をDfOverlapは$\langle pq\gamma \rangle$を計算するクラスで，これらが関与するDfFockmatrix, DfTotalenergy, DfDensityfittingの各オブジェクトからDfEri, DfOverlapのオブジェクトが生成される．

本稿では言及しないが，DfLevelshiftはレベルシフト法，

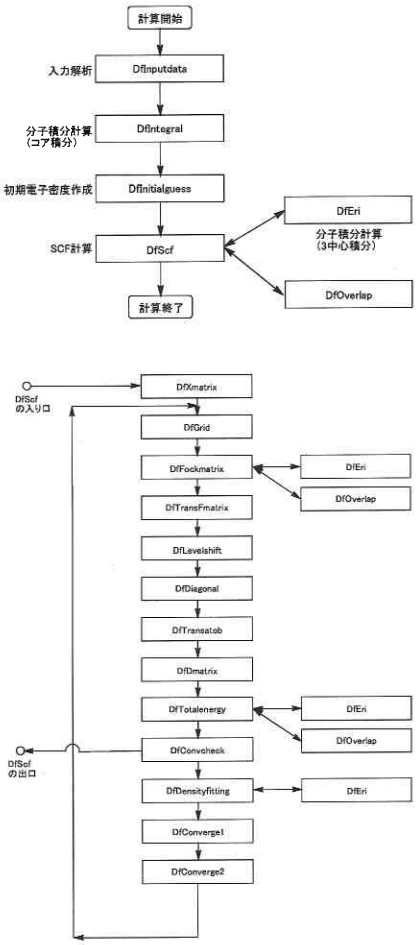

図1　ProteinDFの構造．クラスが示してある．

DfConverge1はダンピング法，DfConverge2はDIIS法を行う．射影演算子法はDfDmatrixの中に，アップデート法はDfEri, DfOverlap関与部に組み込まれている．

図1から明らかなように，分子積分の計算は様々な箇所で必要となる．これらを各必要箇所に組み込んでしまうことは工数的に無駄であり，なにより計算方法を変更すると大掛かりな変更が生じてしまう．そこで，分子積分や行列演算などの計算を行うオブジェクトと，それらを組み合わせてSCF計算を実行するためのシナリオオブジェクトを分離した（図2）．

実は，計算オブジェクトは計算律速のルーチンでもある．オブジェクト指向技術により，仕様を記述するメソッド名はもちろんのこと，引数や戻り値までも変更する必要がないので，プログラムの高速化を考えた場合，図2の分離によりプログラマが手を加える部分は計算クラスだけである．すなわち，計算クラス単体で高速化を行えばよい．高速化手段の差異はシナリオクラスにはなんら影響を与えない．

さらに，様々な計算機アーキテクチャ，通信などのライブラリなどに対応するために，仮想マシンオブジェクトという概念を導入して，プログラムの階層化を行っている．図3はこの概念を使用した並列処理の例である．仮想マシンオブジェクトには各マシン固有の情報を持たせることにより，計算部分からプロセッサに依存する部分を分離した．また，通信ソフトウエアに依存する通信部分も通信クラスにまとめてしまうことで，通信ライブラリ関数そのものを直接呼び出さないようにしている．

このようにして，ProteinDFプログラムはオブジェクト指向を使用して，シナリオ部，計算部，仮想マシン部，通信部に完全に階層化されている[5]．

4. シトクロム c の全電子計算[4]

シトクロム c（口絵図1）は古くから知られた電子伝達タンパク質で，典型的なヘムタンパク質である．ヘム鉄の2価と3価の状態遷移を利用して電子の授受を行う．生体内ではミトコンドリアの呼吸鎖中で膜タンパク質から膜タンパク質へ電子を運ぶ役割を担っている．分子軌道計算に用いた馬心筋シトクロム c は1つの c 型ヘムと104残基のタンパク質部分からなり，ヘムがタンパク質部分のアミノ酸残基との間に2本の配位結合と2本の共有結合で接続されている．原子数は1,738，電子数は6,586である．計算にはProteinDF，基底関数および補助基底関数はValence Doubleを使用し，その総数はそれぞれ9,600，17,578である．

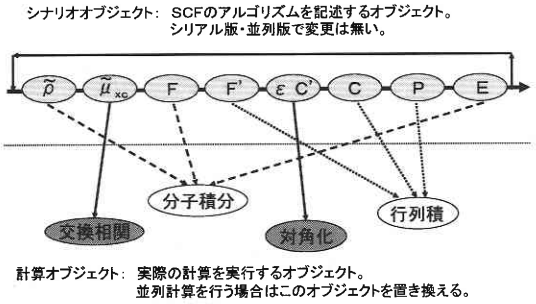

図2　シナリオオブジェクトと計算オブジェクトの分離

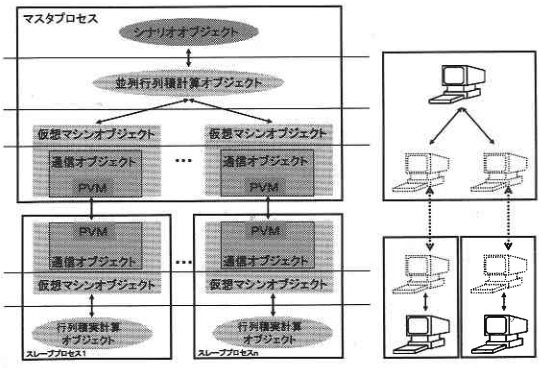

図3　プログラムの階層化による並列化

図4は d^6-low-spin フェロシトクロム c の最高被占有軌道HOMO近傍のKS軌道エネルギー分布と主な占有分子軌道の等値面を分子骨格とともに描いたグラフィックスである．各KS軌道エネルギーは縦棒で描いている．軌道の色は符号に対応している．描いた分子軌道は右からヘム鉄のHOMOである d_{xy}, d_{xz}, d_{yz} に由来する軌道とポルフィリンの a_{2u} に由来する軌道である．ヘム鉄の d 軌道を主成分とする軌道はギャップの中に，ポルフィリンの π 軌道を主成分とする軌道は原子価バンド端に見られるのが特徴である．タンパク質の軌道エネルギーは密に分布しているが，活性中心ヘムに由来する軌道は比較的分散しており，かつヘム単独の軌道の形が保存されているように見える．おそらくこれは金属タンパク質に共通の特徴であろう．

図5のグラフィックスはHOMOを様々な等値面の値で描いたもので，等値面の値は左上から順に，±0.05，±0.005，±0.0005，±0.00005である．Aはスケールを倍にして描いている．HOMOは d_{xy} を主成分とする軌道で，d 成分の総計は85％である．この軌道は面白いことに等値面が±0.0005までほぼ位相を保ちながら，ペプチドの軌道を借りて分子全体に裾を引いている．鉄原子の3d 軌道に比べ，比較にならないほど広がっていて，タンパク質表面でもかなりの値を持っている．シトクロム c は全体が扁平である上にヘムは中心からずれたところにあるので，HOMOはヘムに近い表面でかなりはみ出している．電子移動の担い手であるHOMOのこの特徴は，この軌道がアクセプターの軌道に直接電子を渡している可能性がある．

タンパク質のほとんどの軌道エネルギーは密集している（図4）ので，局在化軌道（Localized Orbital：LO）で表現しても実質的な差違はないが，HOMOのあたりだけは軌道エネルギーが離散的に分布しているので，電子移動などの物性を説明するためにはカノニカル軌道として取り扱わな

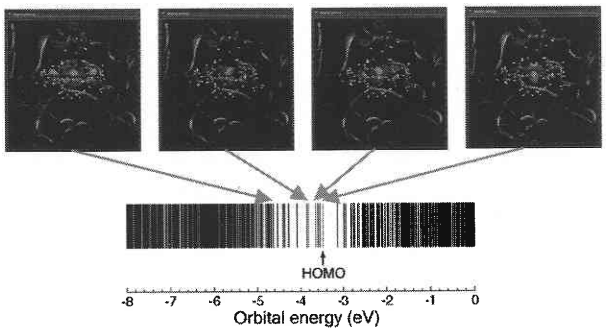

図4　HOMO近傍のKS軌道エネルギー分布と主な占有分子軌道

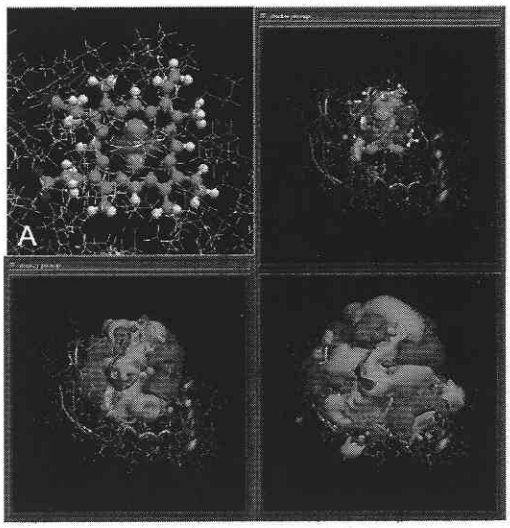

図5　シトクロム c のHOMO

ければならない．

続いて，表1にシトクロム c の全電子計算によるアラニン（ALA）の原子のMulliken電荷を載せた（parm 94のALAの点電荷については5節参照）．全電子計算を行うと，同じアラニンでもタンパク質内の位置によって電荷が異なることが分かる．しかも，その違いは83番目と96番目のアラニンで0.1ほどもあり，電荷の0.1の違いが10 Å遠方で与える影響はほぼ0.1 eV（≒3 kcal/mol）にもなる．生体内で働くタンパク質は，0.1 eV程のわずかな自由エネルギー変化で構造を変化させ，機能を発現しているので，この違いは無視することはできない．この結果からも，タンパク質の全電子計算はタンパク質反応の解析に必要であることが明らかである．

最後に計算量について報告する．シトクロム c の全電子計算はALPHA 21264 667 MHz ワークステーション（WS）1台とALPHA 21164A 533 MHz WS 8台，500 MHz WS 6台の計15台を100 Base-TXのEthernetで接続したWSクラスタ並列システムを使用した．59回のSCF計算を要し，繰り返し計算の1回にかかる実時間は77362秒（21.5時間）であった．

図6は軌道数とSCF 1回の計算にかかった実時間を両対数プロットで示したものである．○，◇，□，△および▽はそれぞれ全実時間，分子積分，交換相関項フィッティング，対角化および行列積を示しており，各傾きは2.4，2.3，1.8，3.3および2.9である．また，金属が入っている分子は各記号を塗りつぶしている．この傾きはそれぞれの計算ルーチンにかかる時間が分子の軌道数の何乗に依存しているかを示している．

このデータと15台のWSによる計算時間実測値から，ProteinDFによるシトクロム c 計算の並列化効率を見積もりたい[5]．当然のことながら，現在1台のWSではシトクロム c の全電子計算を実行できないので，ALPHA 21164A 533 MHz WS 1台で仮に実行できたとして，この時間を予測する．このWS 1台での31残基ペプチド鎖の計算時間実測値をもとに，図6の計算時間の軌道数依存性を利用して，1台での予測時間を算出した．並列化効率は，15台のWSがヘテロクラスタのため，SPECfp_base2000をもとにプロセッサ能力を規格化して求めた．

表1　シトクロム c の全電子計算によるアラニン（ALA）のMulliken電荷とparm94のALAの点電荷

	ALA15	ALA43	ALA51	ALA83	ALA96	ALA101	ALA parm94
N	−0.48	−0.52	−0.48	−0.52	−0.49	−0.47	−0.42
Cα	−0.09	−0.14	−0.11	−0.09	−0.10	−0.10	0.03
C	0.23	0.28	0.21	0.22	0.24	0.25	0.60
O	−0.36	−0.37	−0.34	−0.37	−0.37	−0.39	−0.57
Cβ	−0.61	−0.51	−0.57	−0.59	−0.61	−0.61	−0.18
H	0.41	0.43	0.33	0.38	0.40	0.40	0.27
Hα	0.25	0.20	0.26	0.24	0.24	0.24	0.08
Hβ1	0.26	0.21	0.26	0.25	0.23	0.25	0.06
Hβ2	0.21	0.18	0.24	0.21	0.28	0.24	0.06
Hβ3	0.25	0.26	0.22	0.23	0.25	0.22	0.06
合計	0.05	0.02	0.03	**−0.04**	**0.08**	0.02	**0.00**

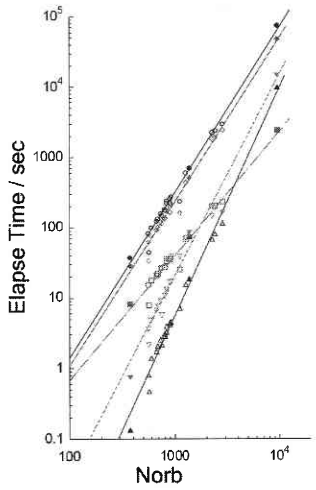

図6　軌道数とSCF1回の計算にかかった実時間の両対数プロット．○，◇，□，△，▽はそれぞれ全実時間，分子積分，交換相関項フィッティング，対角化，行列積である．金属が入っている分子は各記号を塗りつぶしている．

その結果が，表2である．行列積を除き，70％以上の効率が得られた．電子密度フィッティングは，シトクロム c の計算を256 MBしかメモリを持っていないワークステーションに適用するために，1台では行わない特殊なデータ加工処理技術を用いた結果，並列化効率が落ちていることがわかっている．もちろん，今日の計算機資源で実行する際には，その処理は必要ない．交換相関ポテンシャルフィッティングは計算サイズ依存性が小さいため，大規模系では問題ない．行列対角化はこの中で，唯一計算処理のトップループで並列化ができないルーチンではあるが，78％

表2　ProteinDFによるシトクロムc計算の並列化効率の外挿．括弧は行列積を除いた合計の並列化効率．

	1プロセッサでの予測時間（秒）	15プロセッサでの実時間（秒）	並列化効率（%）
全エネルギー計算	273,270	18,779	90.6
Kohn-Sham行列生成	275,767	17,569	97.8
電子密度フィッティング	165,612	14,481	71.2
XCポテンシャルフィッティング	31,891	2,776	71.5
行列対角化	121,170	9,641	78.3
行列積	79,547	14,116	35.1
合計	947,257	77,362	76.3 (85.4)

の効率が出ている．また，大規模系ほど（つまり行列サイズが大きくなるほど）効率がよくなることがわかっているので（データは示さない），こちらも問題はないだろう．

残りの行列積は一見大変問題がある数字のように見える．しかし，解析結果によると，行列積のCPU時間に比べて，100 Base-TXのEthernetによる行列の転送時間があまりにもかかっていることが原因とわかった[5]．行列積の並列化効率を向上するためには，高速なネットワークが必須である．しかし，今日WSクラスタにおいてはMyrinetやギガビットEthernetが主流であること，専用並列計算機の場合は特に高速なネットワークで各プロセッサが接続されるのが普通であることを考慮に入れれば，大規模な並列計算は十分効率よく機能すると考えられる．

これらのデータと計算機の能力を考慮に入れると，スーパーコンピュータを使用して，シトクロムcの10倍ほどのサイズのタンパク質が計算ターゲットに入ってくることが十分予想できる．

5. 次世代量子化学計算システム

今でこそコスト高なタンパク質量子化学計算も，今後の計算機の発展を考慮に入れれば，まもなく誰でも簡単に実行できる時代が到来することは明らかである．これは，これまで古典論による解析が主であったタンパク質の研究に一大革命をもたらすであろう．このような時代に先駆けて，ソフトウエアを整備することが急務の課題となっている．

当グループではProteinDFをベースに，1,000残基規模の超大規模密度汎関数計算が可能なソフトウエアとタンパク質の精密な分子動力学計算が可能なソフトウエアを開発することを目標としている．また，100残基規模のタンパク質の全電子計算または数十残基規模の構造最適化計算を実施し，タンパク質の高精度の電子状態や立体構造のデータベースを構築する．これらの目標をカップルさせることによって，次世代量子化学計算システム（図7）として公開し，本ソフトウエア，データベースともに世界標準（デファクトスタンダード）とすることを目指している．

本研究開発が達成されると複雑なタンパク質の電子状態計算が一気に実用化の域に入り，多くのタンパク質の機能

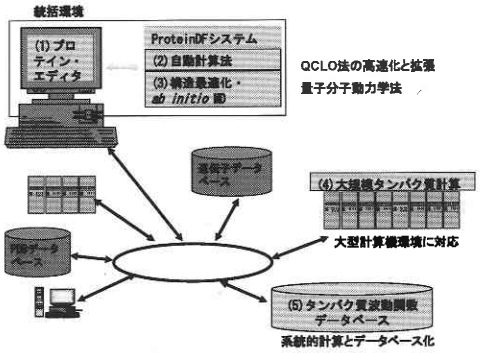

図7　次世代量子化学計算システム概念図

が解明され，生命現象の基礎過程に光があてられるとともに，人工タンパク質の設計が定量的段階に進むと考えられる．本システムは究極のストラクチュラル（構造）バイオインフォマティックス手段であり，基礎研究のみならず産業界においても医薬品，触媒，遺伝子治療，遺伝子改変，環境有害物質の解析等になくてはならないツールに育て上げたいと努力している．また，タンパク質は高効率・高性能の機能素子であることが知られている．特に，光合成反応中心タンパク質やヘムタンパク質などの機能性タンパク質の解析を通して，電子素子，光学素子としての新素材開発，さらにはバイオチップの設計開発への応用に使用されることも期待している．

次世代量子化学計算システムは5つのサブテーマから構成されている．以下にこれらを紹介する．

(1) プロテイン・エディタ

本システムの実行・制御・解析に渡る全機能を統括する環境．(2) から (5) の項目を統合する．これをプロテイン・エディタと仮称している．ユーザによるインタラクティブな制御が必須であるため，グラフィカルなエディタ機能を持つワークベンチ，各種の機能表現に必要なタンパク質構造に有用な大規模分子グラフィックス，およびそのグラフィカル・ユーザ・インターフェースで構成する．プロテイン・エディタは項目毎にサブ環境を持つが，これらはユーザにストレスを与えないよう共通の操作性，共通の表現力をもって統一している．

本システムのデファクトスタンダード化はこのユーザの窓口となるプロテイン･エディタが大きな鍵を握っている．当グループはすでに富士総研と共同でタンパク質全電子半自動計算のためのインターフェース『シナリオ・エディタ』を開発し，ユーザ会で公開した実績がある[11]．次世代量子化学計算システムではさらに，タンパク質の反応解析に役立つツールを作り込み，全機能を容易かつ直感的に操作することができる統括環境へとバージョンアップさせ，ユーザに提供する．

本システムを使用する対象者として，量子化学計算関連の研究者以外にも，生物物理学や生化学の研究者を想定し，プロテイン・エディタを構築している．既存のGUI，グラフィックスはその専門分野に特化したものであり，それが他の分野の研究者による参入を拒む一つの大きな要因となっている．次世代量子化学計算システムのような複数の分野に渡り応用が見込めるシステムでは，このような壁を極力取り払う必要がある．日進月歩である新機能を追加するための拡張性も考慮に入れている．

(2) 自動計算法

タンパク質全電子波動関数の計算には，通常分子の計算とは異なり，専門家が数々の試行錯誤をしなければならない．当グループではLOの一種を使用して安全かつほぼ自動的にタンパク質の全電子計算が達成できる方法を発案した．これにより一般ユーザが容易に複雑なタンパク質の精密な量子化学計算を実行できるようになり，タンパク質計算の普及に役立つと期待できる．

具体的には，タンパク質の全電子計算収束過程法に擬カノニカル局在化軌道（Quasi-Canonical Localized Orbital：QCLO）を導入した新しい方法[12]を自動的に実行する機能を本システムに組み込んだ．現在，その高速計算システムと機能拡張版を開発している．

真の自動化への鍵は，タンパク質の全電子計算収束過程法が握っている．現在の収束過程法は，デフォルトで1残基から3残基，7残基…と機械的に延長して計算を進めるよう設定してある．しかし，タンパク質は複雑な立体構造を持っていため，たとえアミノ酸残基の通し番号が離れていても，三次元空間では近くに配置している場合が多々ある．また，電荷を持ったアミノ酸残基の取り扱いにも注意が必要である．現段階ではユーザ自身が適宜延長シナリオを変更する仕様になっているが，このような情報から安全にタンパク質の全電子計算を達成させる最適な延長シナリオを自動的に提供する方法も研究している．

(3) 構造最適化・*ab initio* MD

現在，タンパク質やDNAなどの生体高分子を原子レベルで取り扱うシミュレーションはAMBERやCHARMMのようなデータベース型の力場を利用したものが主流である．シミュレーション結果の精度に最も大きく影響するのは力場（モデル）であり，シミュレーションの正確さは，採用している力場がどれほど正しいものであるかによる．確かに，これらの力場はアミノ酸残基の種類や原子の種類によって値が異なるよう注意深く作成されているが，問題は，これらのモデルではタンパク質内の位置に依らず，同じアミノ酸残基であればまったく同じパラメータ値を利用することにある．

表1にシトクロムcの全電子計算によるアラニンのMulliken電荷とともにAMBER 94のparmのアラニンの点電荷を示した．両者の定義は異なるので，これらの間で直接値を比較することはできない．しかし，同じアラニンでもタンパク質内の位置によって電荷が異なる全電子計算の結果とは異なり，アミノ酸単位でパラメータを構築するモデルでは，アミノ酸全体の電荷の合計が強制的に整数値（この場合中性なので0）を持たざるを得ないので，電荷の差異などを考慮に入れることは原理的にできない．

そこで，X線構造解析や中性子散乱，多次元NMRなどの実験で得られた構造，およびホモロジーモデリング，古典的な分子動力学法などで決定されたタンパク質のラフな立体構造をもとに，ProteinDFを用いて局所的および大域的な最適化を行い，立体構造の高精度化を行う方法を開発している．また，タンパク質がおかれている環境やタンパク質の構造が時々刻々変化するときのシミュレーション（*ab initio* MD）を実行する方法も開発している．高速計算版用に簡易力場取り込みによる古典分子動力学計算法も併せて提供する．これらはタンパク質の機能解析に必要不可欠な手段である．

(4) 大規模タンパク質計算

これまでの研究結果から，倍精度演算で100,000軌道の全電子計算までは可能であると見積もられている．これは1,000残基規模のタンパク質に相当し，大部分の重要なタンパク質が計算対象に含まれる．そこで，本システムを図4の仕組みを利用して超大型計算機サーバにも対応させる．表3に理論性能40 TFlopsの計算機サーバを用いたときのタンパク質全電子計算SCF一回にかかる時間を見積もった[5]．これを用いて大規模全電子計算の世界レコードをさらに更新したいと考えている．

本研究開発項目の意義は計算サイズに留まらない．大型の機能性タンパク質では，金属などのヘテロ分子を複数持っているのが普通であり，このような複雑な系の量子化学計算はこれまで誰も成功していない．文献12）の拡張したQCLO法が計算成功の鍵を握っている．

(5) タンパク質波動関数データベース

タンパク質は特徴的な機能を持った基本構造と呼ばれる単位から成っている．無数にあると思われたタンパク質の立体構造は，1,000種類程度の基本構造の組み合わせで出来ており，この組み合わせによって機能の多様性が実現されていることがわかってきた．一方で，タンパク質の実験的研究において遺伝子工学的研究は常套手段である．例えば，天然のタンパク質を基に，適当な箇所のアミノ酸残基を置換することによって，より高性能のタンパク質を作成する技術は，創薬研究には欠かせない手法である．

そこで，次世代量子化学計算システムでは，タンパク質の代表的な構造や，同じタンパク質に対して一連の遺伝子工学的な操作がなされた構造をあわせて100種類程度選出

表3　40 TFLOPSコンピュータを用いたときの，シトクロム c および10万軌道タンパク質の1 SCF時間予想

	40TFLOPSでのcytochrome c の予測時間（秒）	40TFLOPSでの10万軌道のタンパク質の予測時間（秒）
全エネルギー計算	7	1,646
Kohn-Sham行列生成	7	1,540
電子密度フィッティング	4	1,270
XCポテンシャルフィッティング	1	76
行列対角化	3	8,804
行列積	2	1,926
合計	24	15,262

し，これらの全電子計算，あるいは構造最適化計算を本システムによって系統的に実行する．ここで，得られたタンパク質の波動関数データや電子密度分布データを収集し，データベースを構築する．このデータベースに特化した計算の結果検索サービス，取得サービス，登録サービスを加え，タンパク質波動関数データベースの完成を目指している．本データベースはタンパク質の実験結果の理解のみならず，タンパク質の新たな物理化学的知見の獲得，新規タンパク質の設計，機能および性能評価に役立つものと期待できる．本データベースのフォーマットそのものもデファクトスタンダードを目指している．

6. むすび

当グループは密度汎関数法による大規模タンパク質の量子化学計算ソフトウエアProteinDFを開発し，これを用いて104残基の金属タンパク質シトクロム c の全電子波動関数計算に世界で初めて成功した．電子相関を取り込んだ100残基の金属タンパク質の全電子計算を行えるソフトウエアは，現在，全世界でこのソフトウエアしか存在しない．当グループの目的はProteinDFを基に，ポストゲノム時代のタンパク質研究，ならびに分子素子の設計などのナノテクノロジーへの応用に必要な機能を追加・拡張し，インフラを整備し，これを汎用レベルにまで完成させ，次世代量子化学計算システムとして公開することである．本システムは以下の5つの研究開発項目（サブグループ），(1) プロテイン・エディタ，(2) 自動計算法，(3) 構造最適化・*ab initio* MD，(4) 大規模タンパク質計算，(5) タンパク質波動関数DBに関する諸研究開発成果を統括したシステムである．本システムはProteinDFを含め計算単位が独立なオブジェクト指向言語C++で構築されるため，各研究開発項目は独立かつ安全に開発することができる．

（2003年3月24日受理）

参考文献

紙面の関係で，最低限の文献のみリストアップした．量子化学全般については1，密度汎関数法全般については2を参照願いたい．

1) 米澤貞次郎他，“三訂量子化学入門（上）（下）”，化学同人（1983）．
2) R. G. Parr, W. Yang, “Density-Functional Theory of Atoms and Molecules”, Oxford University Press (1989).
3) F. Sato *et al.*, *Int. J. Quant. Chem.* **63** (1997) 245.
4) F. Sato *et al*, *Chem. Phys. Lett.* **341** (2001) 645.
5) T. Yoshihiro *et al*, *Chem. Phys. Lett.* **346** (2001) 313.
6) J. C. Slater, *Int. J. Quant. Chem. Symp.* **9** (1975) 7.
7) P. Hohenberg, W. Kohn, *Phys. Rev.* **136** (1964) B 864.
8) W. Kohn, L. J. Sham, *Phys. Rev.* **140** (1965) A 1133.
9) R. Car, M. Parrinello, *Phys. Rev. Lett.* **55** (1985) 2471.
10) B. Stroustrup, “The C++ Programming Language 3 rd Ed.” Addison-Wesley (1997).
11) “ProteinDF System Users Manual Version 0.8 b” (2001).
12) H. Kashiwagi *et al.*, *Mol. Phys.* **101** (2003) 81.