特集

# ポストゲノムプラットフォーム

## Post-Genome Platform ー an Integrated Bioinformatics Solution


**あらまし**

一連のゲノム情報の解読の後，バイオインフォマティクス分野の関心は，得られる大量のゲノム情報をどう活用し，どのような知見を得るかに移っている。このようなポストゲノム時代のバイオインフォマティクスを支え，研究開発を支援する環境として著者らは「ポストゲノムプラットフォーム」と呼ぶ統合ソリューションを提案した。このソリューションは高速・大容量なハードウェアプラットフォーム，これらに最適化されたバイオ基盤ライブラリ，高速な検索エンジン・XML関連ツール，バイオ向けに最適化されたブラウザ，さらにこれらの要素を連携して統合的なバイオ処理の実現を支援する統合ブラウザから構成される。著者らはポストゲノムプラットフォームのコンセプトに基づいて，これらの各要素の開発・整備を行い，実際のバイオ研究の現場で適用検証を行っている。本稿では，このポストゲノムプラットフォームの概要といくつかの構成要素について紹介する。

Abstract

Now that the human genome has been read, researchers in the field of bioinformatics are turning their attention to how they can make full use of the enormous amount of information they have obtained and what kind of knowledge they can acquire from it. To support bioinformatics in the post-genome era, we have proposed an integrated solution called the Post-Genome Platform as a backbone environment for research and development. The Post-Genome Platform consists of a high-speed, high-capacity hardware platform, a basic bioinformatics program library optimized for the hardware platform, a high-speed search engine, XML-related tools, a Web browser optimized for bioinformatics, and a knowledge service accommodator that integrates all of these components and original data or methodologies together so that bioinformatics processing can be executed to create new value. We have developed and improved these components based on the Post-Genome Platform concept and are currently applying them to practical bioscience research for testing. This paper introduces the Post-Genome Platform and describes some of its components.



**奥田　基（おくだ　もとい）**

計算科学技術センター 所属
現在，R&D分野のソリューションの企画・開発に従事。

**松本俊二（まつもと　しゅんじ）**

計算科学技術センター知的システム研究部 所属
現在，知的システムに関する研究開発およびコンサルティングに従事。

**市川眞一（いちかわ　しんいち）**

計算科学技術センターHPCシステム部 所属
現在，HPCソリューションの企画・ビジネス推進に従事。

## ま え が き

一連のゲノム情報の解読の興奮が覚めた後，バイオインフォマティクス分野の関心は，得られる大量のゲノム情報をどう活用し，どのような知見を得るかに移っている。このようなポストゲノム時代のバイオインフォマティクスを支え，研究開発を支援する環境として著者らは「ポストゲノムプラットフォーム」と呼ぶ統合ソリューションを提案した。さらに，これを実現する要素の開発・整備，コンセプトとしての構築を行い，いくつかの要素では実際のバイオ研究の現場で適用検証を行っている。

本稿ではこのポストゲノムプラットフォームについてその概要といくつかの構成要素について紹介を行う。

## ポストゲノムプラットフォームの概要

### ● ポストゲノム時代の現場では

大量のゲノム情報および関連する情報が大量に発生するポストゲノム時代のバイオインフォマティクスの現場では次のような課題が浮き彫りにされている。

(1) バイオ分野の研究者が利用する大量なデータ，各種プログラムは，世界各地にいろいろな形で分散して存在し，日々更新されているという特徴を持ち，そのアクセス，収集，管理が現場の研究者の大きな負担となっている。

(2) これらの大量のデータの中から各研究者が必要とするデータを適切に選択・連携付け，各種のプログラムと連携し解析するためには複雑な処理が必要となっている。

(3) 対象とする研究対象が蛋白質，酵素といった複雑な物質になり，さらにそれらの相互作用の解析を行う必要が生じており，莫大な計算資源を必要としてきている。

これらの問題の解決を目指し，著者らはバイオ統合ソリューションとしてポストゲノムプラットフォームを提案した。

### ● ポストゲノムプラットフォームの提案

ポストゲノムプラットフォームは前述の課題をハードウェアから基盤ライブラリに至るまでの統合的な環境で解決することを目指している。

ポストゲノムプラットフォームの全体構成を**図-1**に示す。ここで示すようにポストゲノムプラットフォームは，

(1) バイオ分野で要求される超高速・大容量の計算を実現するHPCサーバ（スカラSMPサーバPRIMEPOWER，PRIMERGY PCクラスタなど）

(2) これらのサーバに最適化されたバイオ処理の基本となる基盤ライブラリ（相同性検索アプリケーション：BALST，半経験的分子軌道法アプリMOPACなど）

(3) バイオで利用が拡大しているXML形式の大量

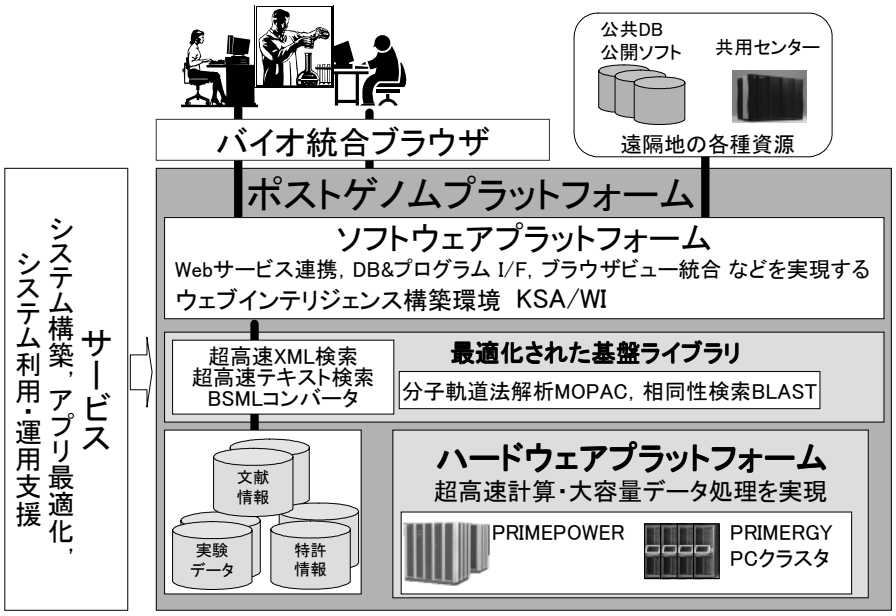



図-1 ポストゲノムプラットフォーム構成
Fig.1-Configuration of post-genome platform.

データの検索，大規模なテキスト検索が可能な高速な検索エンジンおよび各種バイオDBの情報をバイオ分野のXMLであるBSML（Bioinformatics Sequence Mark-up Language）形式に変換しユーザのデータ利用・管理を容易にするコンバータ群

(4) バイオに最適化した表示機能を持つ統合ブラウザGenomic XML Browser

(5) Webサービス機能，各種DBおよびプログラムのインタフェースを持ち，ポストゲノムプラットフォームのほかの要素と連携し，データおよびプログラムの複雑な連携処理を容易に実現するウェブインテリジェンス[1]構築環境KSA/WI（Knowledge Service Accommodator for Web Intelligence）

(6) システム全体の構築支援，バイオアプリの最適化，バイオ向けシステムの利用・運用支援などの各種サービス群

の各要素から構成される。

これらの各要素は単独または統合利用環境の配下で連携して，ポストゲノム時代の研究開発をトータルに支援するソリューションを実現している。

以下の章ではポストゲノムプラットフォームを特徴付けるウェブインテリジェンス構築環境，および基盤ライブラリとなるバイオアプリのプラットフォーム最適化事例について報告する。

## ウェブインテリジェンス構築用環境

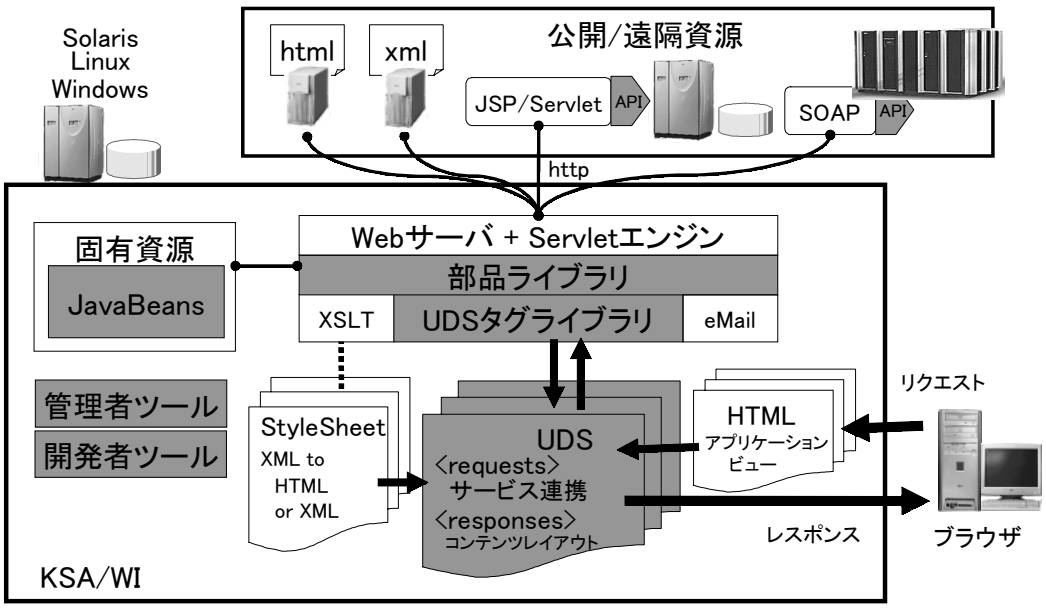
### ● 背景とねらい

ウェブインテリジェンス構築環境KSA/WIは，インターネットを介してアクセスできる公開資源や，企業あるいは研究機関などのイントラネット内の遠隔資源と，研究室や個人で所有しているローカルな固有資源を統一的に連携させ，研究者自身のオリジナルアイディアを付加することで知的ウェブアプリを実現するための枠組みと各種ライブラリを提供する。ここで，資源とは，ネットワークを介して得られる実験データ・解析プログラム・論文・ジャーナル・特許情報などを含む広い概念を指している。

各資源のデータ量・計算量・サイト数は日々増加の一途をたどっており，研究者は，これらを駆使して新しい発見をするべく日夜格闘している。一方，情報技術（IT：Information Technology）提供側としては，優れた解析アルゴリズムなどの開発はもちろん，大量かつ日々変化し続ける分散資源を活用するための基盤作りが大きな課題の一つとなっている。

### ● KSA/WIの構成と主な機能

KSA/WIの構成を**図**-2に示す。KSA/WIの主な構成要素は，UDS（User Defined Service Script）と部品ライブラリである。




KSA/WI: Knowledge Service Accommodator for Web Intelligence,
UDS: User Defined Service Script, JSP: Java Server Pages, SOAP: Simple Object Access Protocol

図-2 KSA/WI構成
Fig.2-System configuration of KSA/WI.

UDSは，XMLのようなタグ形式で記述される一種の簡易言語で，Java ServletやJSP（Java Server Pages）と同様にブラウザを介して呼び出される。アプリケーションの顔（画面）は規定しないので，利用目的に応じてHTMLファイルやクライアントアプリなどにより自由にデザインできる。

部品ライブラリは，一般的なDBMS（OracleやBizSearchなど）へのAPIなどをまとめた汎用部品ライブラリと，バイオ分野でよく用いられる各種公開データベース（DNA配列，蛋白質配列と立体構造，文献情報など）および解析ツール（類似検索，配列比較など）へのAPIをまとめたバイオオプションで構成される。さらに，日々増え続けるツール類を取り込むために部品作成用のAPIが用意されているので，新規プログラムやユーザ固有プログラムも部品化して取り込める。

UDSの主要部分は，リクエスト部とレスポンス部から成る。リクエスト部には，ネットワーク経由およびJavaBeanとして起動できるローカルプログラムに対する検索や計算の要求（リクエスト）を定義する。つまり，HTMLやXMLのサイト（静的なページやServlet/JSPなどにより生成されるページを返すサイト），SOAP（Simple Object Access Protocol）で公開されたリモートプロシジャ，BeanとしてラッピングされたローカルプログラムなどがKSA/WIアプリのアクセス対象となる。一つのUDS内にリクエストは複数定義でき，入出力関係を持つリクエスト（あるリクエストの結果がほかのリクエストの入力データとなる）は一連の処理として順次実行し，それ以外は並行実行するよう自動スケジューリングされる。さらに，バイオ特有の膨大なデータ交換に対応するために，瞬間的に巨大なデータがメモリ上に展開されないよう工夫されている。また，メール通知を指定しておくと，結果を格納したサイトのURLがメール通知されるので，バイオ特有の長時間ジョブ（数時間以上かかる計算など）などがバッチ的に実行できる。

レスポンス部には，対応するリクエストの結果を依頼元に返す形式（レスポンス）を定義する。レスポンスはリクエストに対応して複数定義でき，要求元のブラウザ上で複数のリクエスト結果を一画面にまとめて表示することができる。これにより，異なるサイトの検索結果を並べて比較検討するようなサービスが実現できる。結果をXMLで返してくるリクエストに対しては，スタイルシートでデータ変換することでHTMLファイルとしてブラウザに返したり，ほかのリクエストへの入力にしたりするような使い方ができる。バイオ分野でもXMLサイトは増加しつつあり，スタイルシートを用いたデータ変換により連携の範囲が広がる。

**● 適用例**

バイオ研究者が利用する文献サイトとしてPubMed（米国NCBIのEntrezサービスの一つ）がよく知られている。このサイトへは，日々公開される論文に関する情報がアップされるので，研究者は毎日のように新着情報をチェックしなければならない。必要だが煩わしい作業ではある。UDSと部品ライブラリを組み合わせることにより，ユーザ認証，Entrezへの検索要求，検索結果（XML）からの情報抽出，ユーザごとの新着情報の抽出，スタイルシートによるデータ変換，指定アドレスへのメール通知といった新しいサービスが実現される。これにより，アプリケーションの利用者は，自分の関心ある新着情報だけを自動的に検索してメール通知を受けることができる。UDSの応用としては非常に簡単だが，研究者にとって大変便利なサービスと言えよう。

そのほか，遺伝統計解析プログラムを組み合わせた疾患関連遺伝子発見パイプライン（徳島大学殿），多生物種の類似性に基づく遺伝子相互作用データを活用した推論指向遺伝子関連性探索（理研GSC殿との共同研究）など，先進分野での利用が既に始まっている。

## 最適化ライブラリの実現

ここではポストゲノムプラットフォームの基盤ライブラリとして整備を進めるMOPAC2002のプラットフォーム最適化の例として，スカラ並列マシン上の高速化について述べる。MOPAC2002は汎用の半経験的分子軌道パッケージであり，分子構造や化学反応の理論的解析のために開発された。高速アルゴリズムMOZYMEの採用により，数万原子の系の解析が可能であり，ポストゲノム解析において薬品候補物質と蛋白質活性部位との結合の解析などへの利用が期待される。

分子軌道の計算では分子軌道の固有方程式を解くが，方程式の係数行列自身が方程式の解に依存するため，Self-Consistent Field（SCF）の方法[2]と呼ばれる反復計算方法が使われる。分子電子系の量子力学では，分子軌道関数を原子軌道関数の重ね合わせとして表現する方法が採られ，[2] MOPAC2002が採用している半経験的分子軌道法では，分子電子系のHamiltonianは，実験結果から得られた原子のイオン化ポテンシャルや電子間クーロンエネルギーなどから計算する。さらに，各原子近傍に

局在化する分子軌道を初期値として反復することによって，分子中で直接の結合関係にないもしくは近傍ではない原子の間における分子軌道相互作用の計算を，反復の大部分で省くことを可能としている。[2] これによって，蛋白質のように数万の原子から成り立つ大規模分子の計算が現実的なものとなっている。

● **計算処理の並列性**

MOPAC2002の計算の流れ[3]を以下に示す。

(1) 電子積分[2]と初期値としての局所的分子軌道の計算
(2) 原子軌道についての密度行列[2]の計算
(3) 分子系のFock行列[2]の計算
(4) 電子エネルギーの計算
(5) ユニタリー変換による分子軌道の計算
(6) (2)に戻って繰り返す

(3)では，実験からパラメタとして決めてある電子積分などと，(2)の密度行列を用いた積分計算によって，原子軌道表現のFock行列を求める。(5)では，対角化によらずエネルギー極小となる分子軌道を求める。このためにまず，原子軌道表現のFock行列と，分子軌道の原子軌道展開係数を用いて，分子軌道ごとに分子軌道表現のFock行列対角要素や占有分子軌道・非占有分子軌道間の行列要素を計算する。また，エネルギー極小の条件は占有分子軌道と非占有分子軌道の間の行列要素が0であれば十分である。これを満たすユニタリー変換を，分子軌道の組ごとに分子軌道に対して行う。実際は，分子軌道を表す原子軌道の重ね合わせの展開係数に対して行う。(2)では，原子軌道の組おのおのについて，分子軌道展開係数の積をすべての占有分子軌道にわたって加算する。実際には，各分子軌道は限られた原子軌道の重ね合わせであるため，限定された占有分子軌道について加算される。全体として，並列性に富んだ計算であって，以下の並列処理が可能である。

・密度行列：分子軌道単位の並列計算と原子軌道組合せ要素への総和処理
・原子軌道表現Fock行列：原子ないし原子軌道単位の並列計算
・電子エネルギー：原子単位の並列計算と総和処理
・分子軌道表現Fock行列：分子軌道単位の並列計算
・分子軌道ユニタリー変換：分子軌道単位の並列計算

Bacteriorhodopsin 3686原子の系と，Human Immuno-deficiency Virus（HIV）逆転写酵素と阻害剤の複合体15531原子の系おのおのの，MOPAC2002によるSCF計算について，富士通の共有メモリ型スカラ並列機PRIMEPOWER上での逐次実行時間の分布を調べた。これによると，前記計算（2）が24%，（3）が23%，（4）が2%，（5）が45%となっている。

● **性能評価結果**

上記の並列性に基づいてOpenMPによる並列化を行った。対象サブルーチンは密度行列DENSIZ，分子軌道表現Fock行列DIAGG1，分子軌道ユニタリー変換DIAGG2，原子軌道表現Fock行列FOCK2Z，電子エネルギーHELECZである。残る逐次実行部の時間割合は6%程度である。並列化部のうち特に分子軌道ユニタリー変換では，占有分子軌道と非占有分子軌道の両方について並列化し，分子軌道への原子軌道の追加や変換の回転角を変えた再変換による計算負荷のCPU間のアンバランスを避けるために，非占有分子軌道については循環的割当てを行っている。

Bacteriorhodopsin 3686原子の系と，HIV逆転写酵素と阻害剤の複合体15531原子の系（**図-3**）おのおのの，原子座標を固定したエネルギー計算について，富士通PRIMEPOWER850/16CPU（675 MHz）上で並列化後の性能を検証した（**図-4**）。16CPUにてbacteriorhodopsinが143秒，HIV逆転写酵素・阻害剤が1,978秒と高速である。現状では，並列化されていない部分が16CPU時には40%前後あるが，逐次実行の6倍近い性能を達成している。加速倍率のCPU数推移から，前記以外に各所で行われるCPU間の総和計算も高並列時の性能の阻害要因であることが推測される。分子軌道はおのおの少しずつ異なる限られた原子軌道の重ね合わせで表され，計算結果に依存して原子軌道の組が変化していくため，

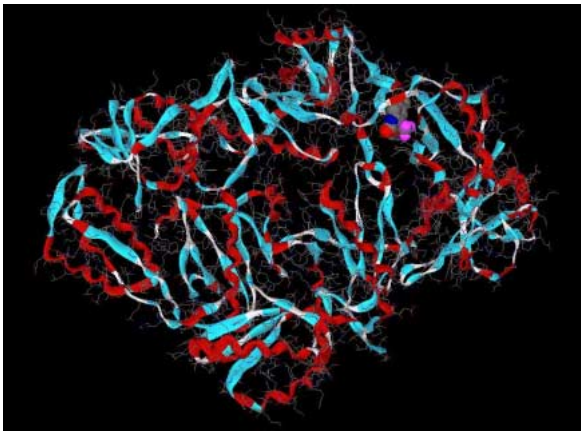
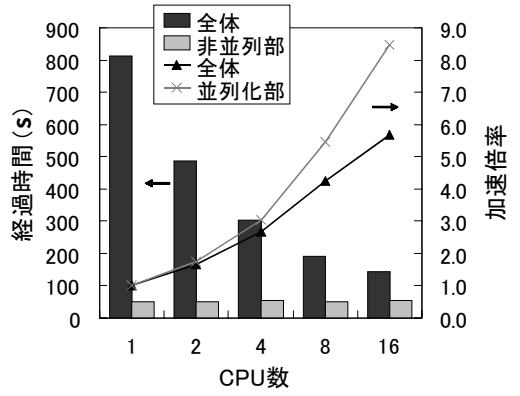
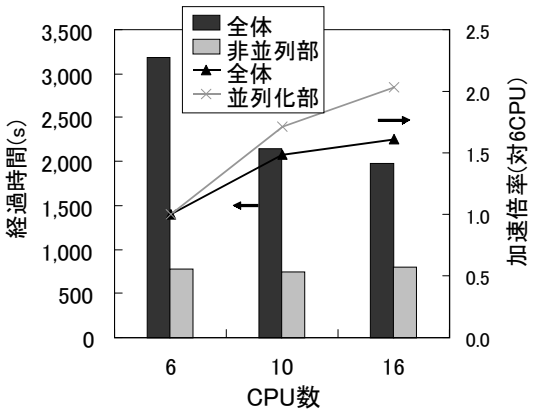


図-3 HIV逆転写酵素と阻害剤の複合体：15531原子
Fig.3-HIV reverse transcriptase (ribbon diagram)-inhibitor complex: 15531 atoms.




（a）bacteriorhodopsin 3686原子




（b）HIV逆転写酵素-阻害剤15531原子

図-4　PRIMEPOWER850（675 MHz）上の性能
Fig.4-Performance on PRIMEPOWER850 (675 MHz).

MOPAC2002ではこれに応じて回帰的にアドレスを決定する圧縮したデータ格納が行われている。総和計算は，この回帰的なアドレス計算を避け並列計算を可能にするために行っている。

### ●　今後について

MOPAC2002とその並列化によって，15531原子もの蛋白質系のSCFエネルギー計算を2,000秒程度で行うことが可能になった。今回の評価では残っていた逐次実行部分について，今後並列化を更に進め，高並列へと性能のスケールアップを目指していきたい。その際にボトルネックになる可能性のある総和計算を回避する並列化を行い，ポストゲノムプラットフォームの基盤ライブラリとして強化を図っていきたい。

## む　す　び

本稿では著者らが提案するポストゲノム時代を支えるバイオソリューション「ポストゲノムプラットフォーム」についてその概要，いくつかの構成要素について説明を行った。

これらの構成要素についてはまだその開発・整備が緒についたばかりであり，今後，機能の強化，基盤ライブラリの整備を続けていく予定である。

このソリューションの適用によりバイオインフォマティック分野の研究開発が更なる発展を見ることを期待している。

### 参 考 文 献

(1)　武田英明：論文特集「Webインテリジェンス」．人工知能学会誌，Vol.17，No.3，p.346-351（2002/5）．

(2)　村井友和：共立物理学講座26　原子・分子の物理学，初版，東京，共立出版，1972年．

(3)　J. J. P. Stewart：“Application of Localized Molecular Orbitals to the Solution of Semiempirical Self-Consistent Field Equations”．*Int. J. of Quantum Chemistry*，Vol.58，p.133-146（1996）．