【別添2】

Ruby 言語による生物化学情報基盤ライブラリの開発
― オープンソースライブラリ BioRuby/ChemRuby ―

1. 背景

1990 年代に開始したヒトゲノム計画は 2003 年にひとまず完了した。ヒトをはじめ様々な生物のゲノムを解析することによって、病気を治すためのメカニズムや生命の歴史を解明していくことができると考えられている。このような解析に必要とされる生物学のデータベースは数千種類あるといわれるが、それぞれ独自のフォーマットを持つことや非常に容量が大きい(GB~TB)ことから、コンピュータを用いた情報処理が必須である。こうした大規模なデータに基づく近年の医学や生物学を根底で支えているのが生物情報学(バイオインフォマティクス)である。ゲノムプロジェクトを契機として、バイオインフォマティクスは医学や生命科学に携わる多くの人に利用される重要な技術となってきたが、この分野の進展が速かったことや個々のニーズが多様であることから、これまで情報科学に縁のなかった実験科学者にも使いやすい環境は十分に整っているとはいえない。そのような状況の中で、ゲノムの実体である DNA 配列やタンパク質のアミノ酸配列の操作、各種データベースや様々な解析プログラムへの対応、といった基盤的な部分を扱うためのライブラリとして、海外では BioPerl, Biopython, BioJava などが開発されてきている。しかし、これら既存のライブラリにも、カバーする領域や言語そのものの使いやすさなど様々な点で問題も残されている。また、国内の様々なデータベース等の有用なリソースが生かせていなかった。

一方で、化学分野でも化合物から薬品まで様々な化学構造のデータベースが作られてきたが、歴史的に商用ソフトウェアの利用が一般的であったため、各社のデータベース間やソフトウェア間での互換性のなさが問題となっていた。近年、ゲノムから明らかにされた遺伝子などの情報と化合物の情報を組み合わせて創薬につなげたい、といった需要が高まっており、化学情報学(ケモインフォマティクス)が注目されてきている。しかし、既存の商用ソフトウェアは GUI による操作が中心で、ハイスループットなバッチ処理といった自動化は困難であった。

そこで、本提案では Ruby 言語を用いたフリーのオープンソースソフトウェアとして、BioRuby ライブラリの拡張と ChemRuby ライブラリの開発を行う。BioRuby では、先行する海外の既存ライブラリが中心課題としてきたゲノムの配列解析だけでなく、遺伝子発現やパスウェイ解析といったポストゲノム時代に求められる新しいタイプのデータ解析に必要な機能の実装を進める。ChemRuby では、様々な化学構造を扱い類似構造検索を行うための機能を実装し、ハイスループットな解析を自動化できることを目指す。さらに、これらのライブラリを普及させるため、使いやすいユーザーインターフェイスの開発、英文と和文によるドキュメントの整備、安定動作を保証するためにユニットテストの増強に取り組む。

2. 目的

BioRuby はすでに 4 年間にわたって開発を進めてきており、基本的な機能はすでに実

装済みであったが、ドキュメントの整備不足や開発にあたる人的資源の不足から十分に普及しているとはいえない。そのため BioRuby の機能開発に加え、英文と和文によるチュートリアルの作成、ライブラリの API リファレンスの整備、信頼性向上のためのユニットテストの実装を行なう。さらに、ChemRuby の新規開発によって、生物化学情報学へと対象領域の拡充を行なう。開発した成果については国際会議等での発表を行なうとともに、講習会などを開催し普及に努める。

3. 開発の内容

以下の開発内容をまとめ、BioRuby 1.0 と ChemRuby 1.0 をリリースした。これらのライブラリは、Ruby 言語が実行可能な環境であれば OS によらず利用できる。

機能開発

塩基・アミノ酸配列クラスの改良、BLAST サポートの改良、HMMER レポートクラスの改良、GenBank/EMBL クラスの改良、fastacmd への対応、siRNA デザインツールの採用、コマンド入出力フレームワークの改善、KGML パーザ、ColorScheme の採用、アラインメントクラスの改良、PDB クラスの改良、制限酵素への対応、EMBOSS USA への対応などの機能拡張・改善と、その他数多くのバグフィックスを行った。また、Ruby の autoload 機能を採用することにより、従来は1秒近くかかっていたライブラリの読み込みを約 40 倍高速化し 0.02 秒まで短縮することができた。

さらに、これらの成果を Ruby プログラマ以外のユーザにも広く使いやすいものとするため、コマンドラインインターフェイス BioRuby シェルを作成した。またシェル上のオブジェクトを Ruby on Rails を利用してウェブブラウザで表示する機能を実装した。(図1)

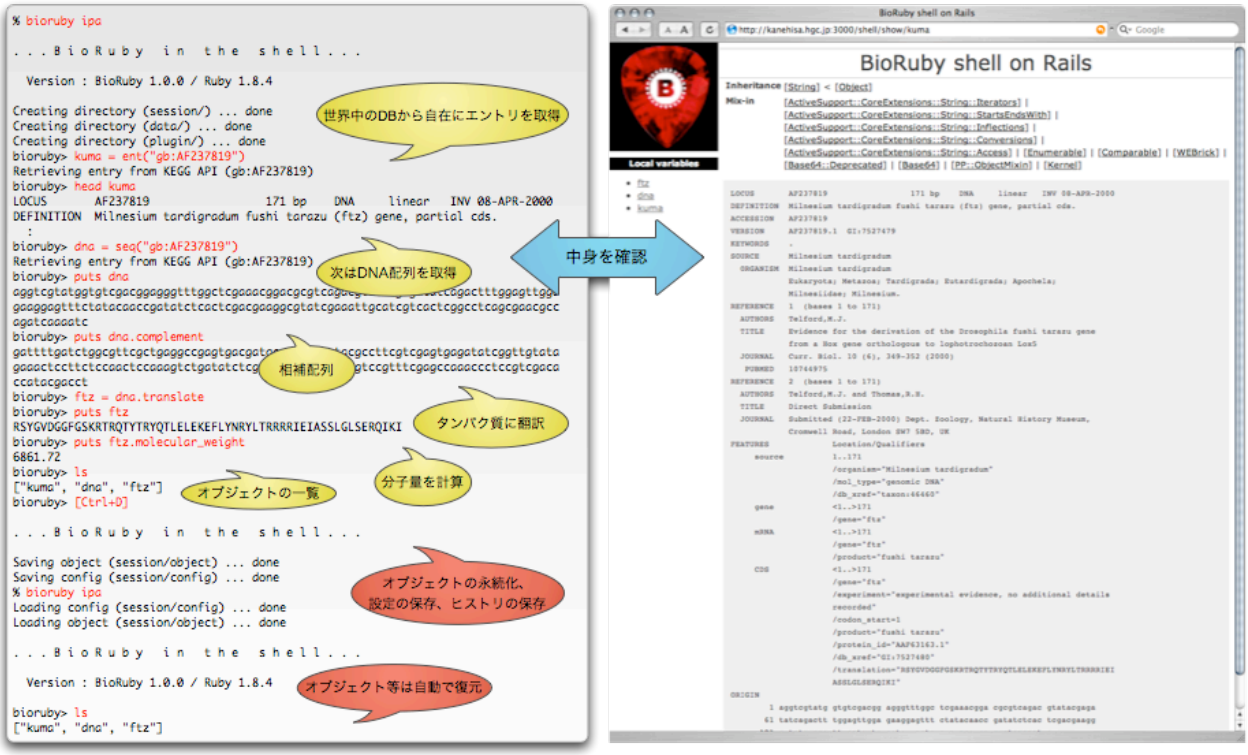

図1 BioRuby シェルと Ruby on Rails を利用したシェルブラウザ

ドキュメントの整備

和文チュートリアル文書の充実をはかり分量を倍増するとともに、海外への普及を促進するために英文チュートリアル文書を作成した。また、これまでライブラリの各ファイルに分散していた API のドキュメントを Ruby 言語で主流になりつつある RDoc フォーマットに変更し書き足すことにより、ウェブサイトで閲覧できるようにした。(図2)

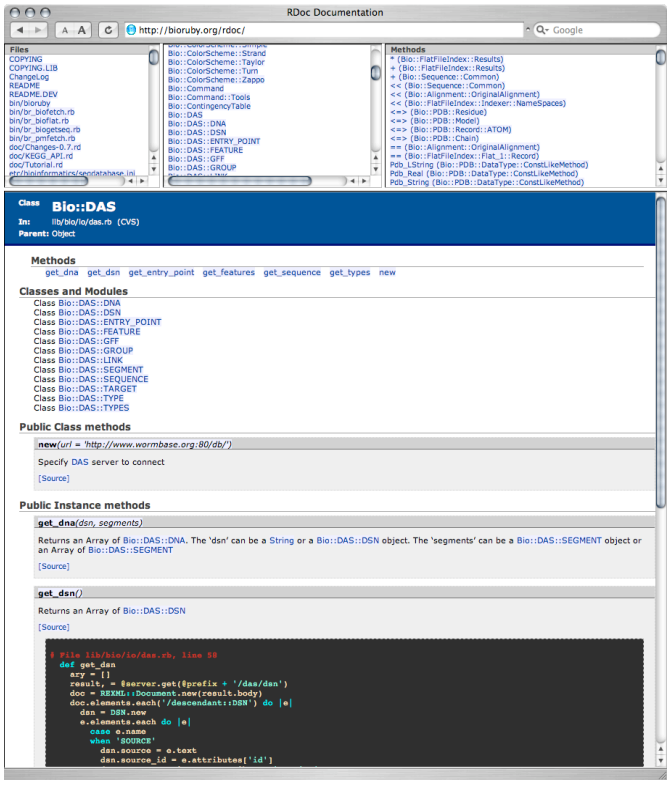

図2 RDoc による API ドキュメント
http://bioruby.org/rdoc/

図3 Test::Unit によるテストの実行

```
k@sw:/usr/local/src/ruby/bioruby/bioruby-1.0.0% ruby install.rb test
install.rb: entering test phase...
Running all tests in test...
Loaded suite test
Started
.....
Finished in 32.294052 seconds.

929 tests, 1253 assertions, 0 failures, 0 errors
install.rb: test done.
k@sw:/usr/local/src/ruby/bioruby/bioruby-1.0.0%
```

ユニットテストの開発

科学分野で利用されるライブラリとしての安定性を確保し、開発に伴う変更に対してロバストなものとするため、Ruby 標準添付の Test::Unit フレームワークを利用した約 1000 種類のユニットテストを作成した。(図3)

ChemRuby の開発

生物分野をカバーする BioRuby に加え、新規開発ライブラリとして、化学分野をカバーする ChemRuby を開発した。ChemRuby では各種商用ソフトウェアの様々な化学構造フォーマットの読み込みに対応し、高速に類似構造を検索をする機能を実装した。既存の商用ソフトウェアは GUI を用いているためバッチ処理が困難であったが、ChemRuby により結果のビジュアライズまでをプログラム化することが可能となった。(図4)

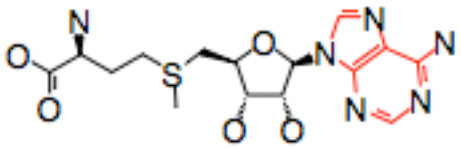

図4 ATP 分子とアデニン分子の共通部分構造を検索し、一致部分を着色した例

4. 従来の技術（または機能）との相違

BioRuby は、他言語用の類似ソフトウェアと同等の機能を持ちながら、カバーする領域は配列解析にとどまらず、ポストゲノム時代の解析にも活用できる。特に、新規開発したシェルは他にはないもので、誰にでも利用しやすい環境を提供できた。ChemRuby については、化学分野でこれまでフリーの類似ライブラリがなかったため、自作プログラムへの組み込みからハイスループットなパイプライン処理まで利用価値は高い。また、これらのインストール方法は、一般的な Ruby ライブラリに準拠しており、Ruby 自体の標準添付ライブラリが充実しているため、他言語用の類似ソフトウェアと比較して導入が極めて容易である。

5. 期待される効果

BioRuby/ChemRuby が利用される分野としては、第一に大学などの研究機関における研究活動や教育実習などが挙げられる。研究教育機関においては、予算面からソフトウェアがフリーであること、また科学的な透明性の観点からオープンソースであることが重要である。今後、バイオインフォマティクス・ケモインフォマティクスの学部教育が広まるとともに、BioRuby/ChemRuby を使用することで、商用ソフトウェアと比較して学生数×年次分のコストダウンにつながることが期待される。また、Ruby 言語用のライブラリであるため、ウェブ CGI の開発や自作プログラムの一部に組み込んだ利用も容易であり、すでに内外の研究者によって使用されている。第二には製薬を中心とした企業での利用が想定される。これらの企業では BioRuby と ChemRuby を活用して、インハウスのデータベースや解析パイプラインを開発することにより、コストを抑えながら社内のデータ流出を防ぐことが可能となるため、今後の需要が見込まれる。

6. 普及（または活用）の見通し

BioRuby の過去2年間での累計ダウンロード数は 14,000 回であり、最新版のダウンロード状況から推定した現時点での利用者数は全世界で 300 人以上と考えられる。このうち国内の利用者は約 80 人程度、海外からの利用者は約 150 人、取得元が IP アドレスのため国別不明のものが 70 人程度となっている。所属機関では ac や edu ドメインなどアカデミックが多い傾向にある。これまでダウンロード数はリリース毎に増加してきたが、今回の未踏の成果をもとに学会や講習会などでの普及活動を継続することによって、広く普及できると考えられる。ChemRuby については、今後認知度を高めていく必要があるが、その有用性からすでに一部の企業でも使われている。

7. 開発者名（所属）

片山俊明（東京大学医科学研究所・ヒトゲノム解析センター）
中尾光輝（産業技術総合研究所・生命情報科学研究センター）
後藤直久（大阪大学微生物病研究所・遺伝情報実験センター）
田中伸也（京都大学化学研究所・バイオインフォマティクスセンター）

（参考）開発者URL

http://bioruby.org/ と http://chemruby.org/ を参照。