家畜の遺伝子地図作成のためのソフトウエア秘書を目指して

農林水産省畜産試験場育種部

和田康彦

ywada@niai.affrc.go.jp

要約

農林水産省畜産試験場では農林水産省家畜衛生試験場、STAFF研究所、道県の畜産試験場等と共同で豚の遺伝子地図作成のプロジェクトを進めている。この中で、論文発表済みの遺伝子地図情報の検索を目的とした家畜ゲノムデータベース(http://ws4.niai.affrc.go.jp/)、連鎖解析のためのデータ処理の一元化を目指したオンライン連鎖解析データベース、およびそれらのクライアントであるgRanchを開発してきた。しかしながら、gRanchはパソコン上で動作できず、また、分散ネットワークシステムとの相性も良くない。そこで、Java言語上のオブジェクトブローカーであるHORBを用いた連鎖解析のためのソフトウエア秘書を目指して、既存のシステムをJava言語とHORBで書き換える作業をしている。本論文ではその現状を述べるとともに、他のセクションを含めた遺伝子地図作成のためのソフトウエア秘書システムの全体設計を行なう。

Assistant software for gene mapping

Departmenet of Animal Breeding and Genetics, National Institute of Animal Industry

Yasuhiko Wada

Abstract

We are currently mapping the porcine genetic markers for linkage analysis to elucidate genes or genetic regions responsible for phenotypes under the collaboration of the National Institute of Animal Industry, National Institute of Animal Health and STAFF institute. For this analysis, the online linkage analysis system have been developed. We have also developed a gene mapping workbench named gRanch for data registration and computing lod scores. However, the gRanch can not be used on personal computers and it is difficult for collaboration of multiple CPUs on these systems. Therefore, we have developed a new system for assistance of linkage analysis using the Java language and the HORB, which is an object broker for the Java language. This system includes an automatic correction of allele types, display of alleles transmission in families, normal distribution test of traits data.

はじめに

農林水産省畜産試験場では農林水産省家畜衛生試験場、STAFF 研究所、道県の畜産試験場等と共同で豚の遺伝子地図作成のプロジェクトを進めている。この中で、論文発表済みの遺伝子地図情報の検索を目的とした家畜ゲノムデータベース (http://ws4.niai.affrc.go.jp/)[5]、連鎖解析のためのデータ処理の一元化を目指したオンライン連鎖解析データベース [6]、およびそれらのクライアントである gRanch[7] を開発してきた。また、家畜ゲノムデータベースの一部では Java Applet を用いたビジュアリゼーションを試みている [9]。

これらのシステムはわれわれの連鎖地図作成プロジェクトの中で日常的に利用されているだけでなく、家畜ゲノムデータベースなどインターネット上で公開されたシステムには予想以上のアクセスをいただいている。しかしながら、表形式データベース管理システムと WWW および Java applet の技術を用いた現在のシステムでは次のような問題点が顕著になってきた。たとえば、遺伝子関連の情報の構造が複雑で表形式データベースとの親和性がよくない、入力されたデータに誤りやデータ間の不一致が多い、連鎖地図作成作業を自動化するために必要な遺伝学、分子生物学的背景知識をとりこんだプログラミングが非常にむずかしいなどの問題がある。

これらの問題点を解決するための第 1 歩としてシステムを Java 言語 [4] と HORB[2] で全面的に書き換える作業を進めている。この組合せを選んだ最大の理由は HORB が分散オブジェクトの実行環境として最も優れていると著者が評価したからである。HORB は Java 言語ベースであるために既存の C 言語で書かれたプログラムを移植しやすく、JDBC が制定されているためにデータベース処理を DBMS から独立して記述できる。また、HORB はプログラマに過度の制限を強いることなく自由なプログラミングが可能であり、にもかかわらず非常に性能が良い。また、JDK1.0 ベースのマシンでも実行可能であり、OS とは別個の独自の分散アクセス制御ができるのも魅力である。さらに HORB を用いることで重い処理をサーバー側で実行できるので、小さなリソースしか持たないパソコンでもクライアントとして利用できる。

新システムの一部をすでに連鎖解析のためのアシスタントソフトウエア [8] としてデモ公開しているので、本論文ではその概要について述べるとともに、他のセクションを含めた遺伝子地図作成のためのソフトウエア秘書システムの全体設計をし、実装にあたっての問題点について検討する。

連鎖解析のためのアシスタントソフトウエア

システムの概要を Figure 1 に示したが、現状のシステムはクライアントにダウンロードされる Java Applet、サーバー側の WWW サーバー、HORB サーバー、DBMS の 4 つの部分から構成されている。プログラムは JDK1.1 と HORB1.3 に準拠して、基本部分を Java 言語で記述している。そして、ユーザーインターフェースに関わる部分のみを Java Applet とし他の部分は HORB 上の Java プログラムで処理することで、リソースの少ないパソコンでも軽快に利用できるように配慮している。また、開発期間を短縮するために、既存の他言語のプログラムも使用可能なものは積極的に利用している。さらに、今後、帰納推論や仮説推論などの人工知能的な技術を利用することを考慮して、ユーザーの行動をログとしてデータベース上に格納するようにしている。

ユーザ認証機構としては端末の IP アドレス、ID、パスワードの組をチェックし、ユーザーの所属するプロジェクトとユーザーの権限をデータベースから入手する。データは全てプロジェクトごとにデータベースに格納され、ユーザーは自分の所属するプロジェクト以外のデータを利用することはできない。

システムの機能としては、血統データの入力と修正、アリルデータの入力と修正、指定されたマーカーに対する遺伝子頻度と遺伝子型頻度の算出、親子間のアリルタイプのチェックと自動修正、連鎖地図上でのダブルリコンビナントのチェック、全マーカーに対する lod score 分析、 形質値に対する基礎統計量の算出などが現時点で実装されている。今後、指定されたマーカーに対する連鎖不平衡の検定、マーカー連鎖地図の自動作成、QTL 解析の自動化などの機能の追加を予定している。

この中で lod score 分析はシステムに対する負荷が大きいために、その日にデータの修正があって、かつアリルタイプに矛盾がない場合に限り、夜間に自動的に実行される。ユーザーは次の日にその結果をクライアント上に表示できるようになっている。

連鎖解析は Java プログラムから CRI-MAP 2.4[1] を外部プロセスとして起動し、その出力ファイルを Java プログラム側で処理している。DBMS としては現バージョンでは postgreSQL6.2.1[3] を利用し

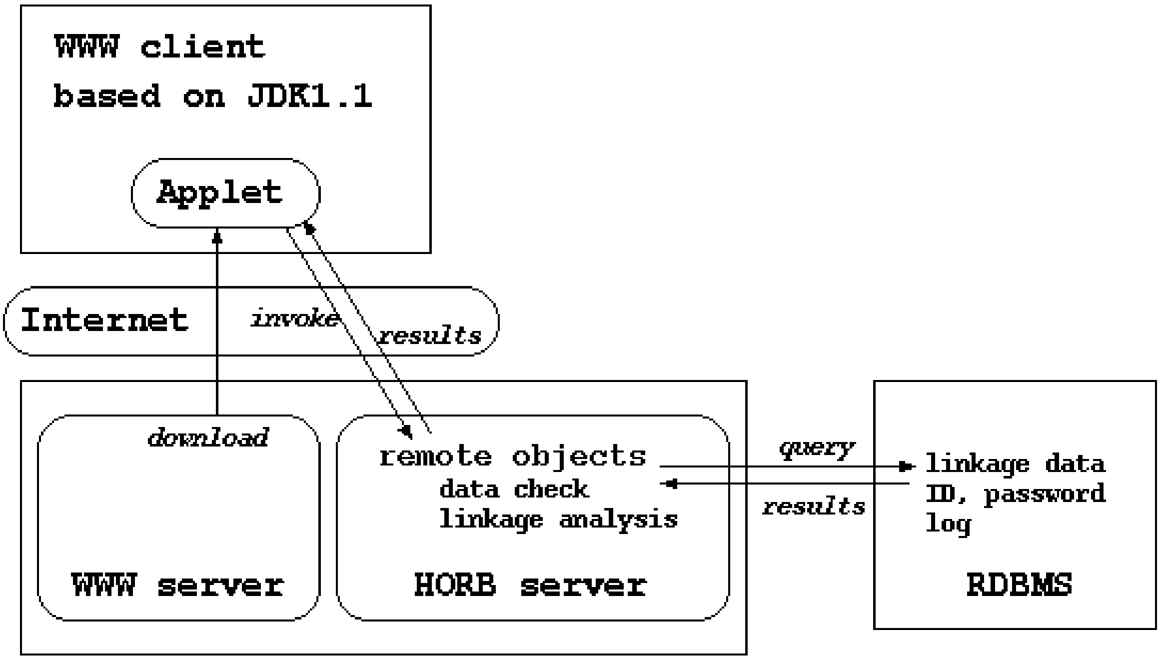

Figure 1: System overview of assistant software for linkage analysis

ている。本システムを構成している Java 言語クラスについて述べると、アプレットへダウンロードするクラスが 22 クラス（2748 行）、horbc でコンパイルしたリモート実行可能なクラスが 16 クラス（2797 行）、リモート実行可能なクラスからサーバー上で呼ばれたり、夜間に cron で実行されるクラスが 15 クラス（1454 行）となっている。すなわち半数以上のコードがクライアント側ではなくサーバー側で実行されているので、その分、クライアント側にかかる負荷を小さくすることが可能になっている。

なお、本プログラムのデモ版は http://genome.niai.affrc.go.jp/linkagent/から試用できる。

システムの全体設計

ここまでで述べてきたシステムは遺伝子地図作成のための秘書ソフトウエアの連鎖解析を担当するセクションであったが、全体のシステムは Figure 2 に示すように多くのセクションから構成される。連鎖解析セクションはこの図では左上端に相当し、その他、マーカーセクション、cDNA セクション、BAC ライブラリセクション、RH パネルセクション、塩基配列セクション、文献セクションなどがあり、それぞれのセクションにはデータベースと、Java 言語と HORB を用いて開発するオブジェクト群を配置する予定である。Figure 2 では紙面の都合上、各セクションに 1 つのオブジェクト群しか書かれていないが、実際には役割別に複数のオブジェクト群が配置される。

また、タンパク質に関する知識ベース、分子生物学一般に関する知識ベース、一般常識に関する知識ベースを構築し、これらとのインターフェースをとるオブジェクト群も Java 言語で開発し、HORB を通して他のセクションのオブジェクト群と共同して処理をすることを考えている。

さらに、各ユーザーに専用の秘書オブジェクトを割り当てて、ユーザーに対する助言、忠告などができるようにしたい。この秘書オブジェクトはマネージャーオブジェクトからの指令により、夜間に当該ユーザーに関連するセクションのオブジェクト群を起動して、現状を分析し、今後ユーザーがとるべき方策について検討する。その結果を保持しておき、次にユーザーがアクセスした時にそれを提示する。もちろん、ユーザーが直接に秘書オブジェクトに尋ねることもできるようにしなければならない。

ユーザーから見た利用イメージとしては、Figure 1 に示した連鎖解析セクションと同じく、WWW クライアントに WWW サーバーから自分用の秘書オブジェクトのユーザーインターフェース部分をダウンロードして、必要な処理を指示したり、自分用の秘書オブジェクトからのメッセージを表示すると

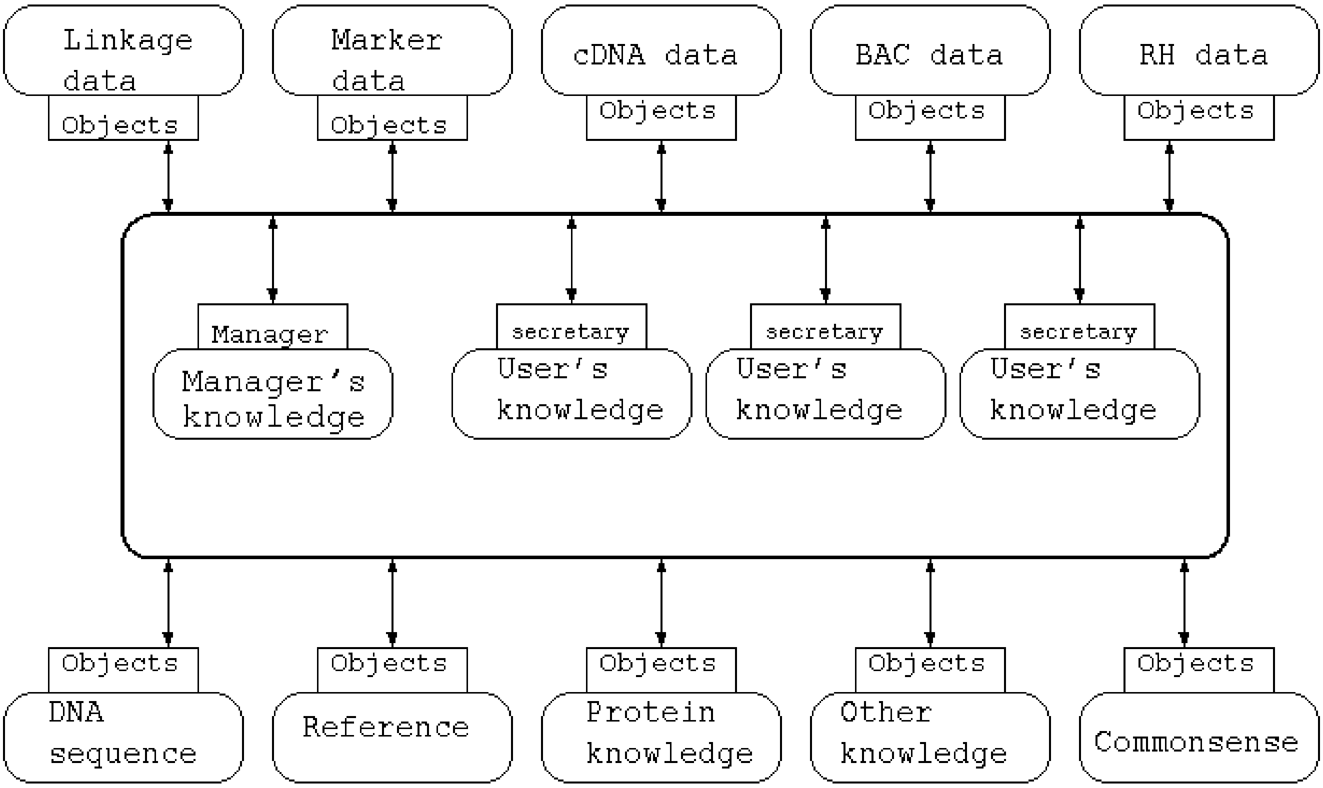

Figure 2: System overview of assistant software for gene mapping

いった使い方になる。

図には特に示さなかったが、従来から提供している家畜ゲノムデータベースの機能もこのシステムの中に吸収したいと考えている。データ入力者用の秘書オブジェクトを作成し、公開データベースについてはWWWサーバーからAppletをダウンロードする方式で、連鎖解析セクションと同様の手法で対処できると考えている。

ユーザー用の秘書オブジェクトのクラスとしては、スーパーユーザー用、ゲノム研究者用、データ入力者用などを用意し、各ユーザーごとにオブジェクトを生成することになるが、各ユーザーのための知識ベースを随時蓄積し、その中から帰納推論機構を用いて各ユーザーの処理スタイルのようなものを学習できるようにすれば、秘書ソフトとしての利用価値が増大すると思われる。

実装上の問題点

実装上まず決めなくてはならないのはオブジェクト間の通信方法である。現状の連鎖解析のためのアシスタントソフトウエアではアプレットからHORBを用いてサーバーオブジェクトを起動し、そのサーバーオブジェクトがさらに必要な別のサーバーオブジェクトを起動するという方法をとっている。この場合、複数のユーザー秘書からの要求があると複数のセクションオブジェクトが生成されるので、リソースを消費するだけでなく、複数のセクションオブジェクトの行動をどのようにコントロールするのか、という問題がある。

セクションオブジェクトをデーモンオブジェクトとして起動すると、上記の問題は解消されるが、メモリ上に全てのセクションオブジェクトが存在するのはリソースの浪費につながるので、不要なオブジェクトは永続オブジェクトとしてファイル上に保存して、必要に応じてロードする方が良いだろう。

DNA配列セクションのように大規模なセクションでは、セクション内での負荷分散を考慮する必要がある。その時にはエージェント的な機能が有効である。具体的には著者が考案したエージェントもどき(http://ws4.niai.affrc.go.jp/ywada/horb/agent.html)の手法を応用することになろう。これは現バー

ジョンのHORBにエージェント機能がないための苦肉の策ではあるが、実用上ほぼ十分な機能と自由度を持っている。HORB Openで開発が開始されたObject Traderはエージェントシステムではないが、分散システムの中で特定のオブジェクトをその機能から検索することが可能であり、セクション内での負化分散を行なう時に利用できるかもしれない。

知識ベースの構築方法と推論、計画のための手法については、現時点では未知数の部分が多い。知識ベースの構築方法としては大きくわけて3種類が考えられる。それらはJava言語で独自に開発する方法、Quixote[10]などの専用の知識ベース管理システムを利用する方法、Common Lispのようなより人工知能的な処理に適した言語でシステムを構築する方法である。

Java言語で独自に開発する方法はシステム構築の自由度が高い反面、プログラミングコストがかかる。また、Java言語は処理速度が遅い、処理系が枯れていない、多重継承ができない、動的型指定ができないといった問題点がある。すでに存在している商用のオブジェクト指向DBMSを利用すれば、これらの欠点の一部は解消されるが、Java言語そのものの束縛からは逃れることができない。

専用の知識ベース管理システムを利用する場合は、一見便利なようであるが、ユーザーインタフェースを書き換える必要がある。特にJava言語以外の言語で作成されたシステムは言語間のインタフェースを別途作成しなければならない。また、研究用のシステムでは一部の手法のみしか実装していない場合が多く、特に推論と計画を同じ知識ベースで利用できる例はほとんどないだろう。

従来のソフトでは、マーカーの連鎖地図を作成するような簡単な場合でも、入力データにミスがある場合は信頼性のない結果を算出したり、連鎖地図を作成できなくなる場合もある。そのような場合にも、的確に問題点を抽出し、ユーザーに助言することが必要になってくる。このような処理をする手法として、簡単なプロダクションシステムを構築することも考えられるが、実用レベルを目指すとすると制約論理や重みつき仮説推論、帰納推論などを考慮しなければいけないかもしれない。しかしながら、これらの手法はまだ実験段階であり、実用レベルで利用できそうなシステムが見当たらないのが現状である。

ただ、どのような手法をとるにしても、手続き的な記述方式では、システムが拡張されるにつれて複雑度が指数関数的に増加して、このシステムそのものが破綻しかねない。どうしても、知識処理的な部分は、宣言的な記述とそのインタプリタあるいはコンパイラという組合せにしなければならないと考えている。

これからの開発予定

この論文でとりあげたシステムは基本的には農林水産省畜産試験場を中心とする家畜ゲノム解析研究グループの研究推進のためのツールとして、実際に利用されることを想定しているものである。従って、現時点で稼働しているシステムを一度に新システムに移行するのはリスクが大きいので、ユーザーに負担のかからない形での移行スケジュールを考える必要がある。また、移行のためのプログラミングコストを小さくするために、セクションによっては現システムの上にJava言語のインタフェースをかぶせる形をとる必要もあろう。

しかしながら、本システムはGenome informatics研究のためのワークベンチとしての利用価値も高い。そのため、新たな手法や新たな実装方法を試みることを恐れてはいけないだろう。実用とInformatics研究とのバランスをいかに取るかが重要になってくる。

これからの開発予定としては、まず、連鎖解析セクションの本格運用に入るとともに、簡単なプロダクションシステム的な機能を備えた連鎖解析ユーザー用秘書クラスのプロトタイプを実装したい。

その後、ここでの経験を生かして、文献セクション、マーカーセクション、BACライブラリセクションの設計、実装を行なう予定である。

References

[1] Green, P., Falls, K. and Crooks, S., “Documentation for CRI-MAP, version 2.4 (3/26/90)”

[2] Hirano, S. “HORB: Distributed Execution of Java Programs”, Worldwide Computing and Its Applications (Springer Lecture Notes in Computer Science 1274) pp.29-42, 1997.

[3] PostgreSQL Organization, “PostgreSQL”, http://www.postgreSQL.org/.

[4] Sun Microsystems, “The Java language specification” 1996.

[5] Wada, Y. and Yasue, H., “Development of an animal genome database” *CABIOS*, Vol. 12, No. 3, p.231-235, 1996.

[6] Wada, Y. and Yasue, H., “Information System for Gene Mapping in Livestock” *Proc. of the 8th AAAP Animal Science Congress*, Vol.1, 544-553, 1996.

[7] Wada, Y., Inoue, K., Ohga, K. and Yasue, H., “Software tool for gene mapping: gRanch” *CABIOS*, Vol. 13, No. 3, p.323-324, 1997.

[8] Wada, Y. and Yasue, H., “Assistant Software for Linkage Analysis” *Genome Informatics 1997*, p.354-355, 1997.

[9] 和田康彦、安江 博, “Java Appletを用いた遺伝子地図情報に関するビジュアリゼーションの試み” コンピュータサイエンス, Vol. 4, No. 2, p.111-115, 1997.

[10] Yokota, K., Kunishima, T. and Nakanishi, H., “Hypothetical Query Processing for Workflow Management” *Proc. the seminar on Logical Databases and the Meaning of Change, (UPMAIL Technical Report No. 129, Uppsala University)*, p.44-49, Dagstuhl, Germany, Sep. 23-27, 1996