共同研究全般にわたるデータベース構築・管理
４の２（地域分）
（フェーズⅠ）①リレーショナルデータベース技術の適用性検討②新システム技術との融合
（フェーズⅠ） ① リレーショナルデータベース技術の適用性検討② 基盤技術の応用
（フェーズⅠ） 遺伝子データ解析技術に関する研究
（フェーズⅡ）遺伝子データ解析技術に関する研究
（フェーズⅡ）効率的なプロテオーム解析のための総合的研究環境を実現するシステム構築に関する研究
（フェーズⅡ）大規模データの分析・活用技術の研究タンパク質や遺伝子のデータなどの大量の蓄積データから有益な情報を効率よく抽出するためのシステムを構築する。

新日鉄ソリューションズ株式会社（共同研究員） 小野 祥正、嶋本 公徳、小橋 由佳

[目的と概要]

薬剤ターゲットや遺伝子診断に役立つ遺伝子の機能解析研究をＩＴ的に支援するための技術研究開発が目的である。ここで、研究開発をした技術をもとに主に国内製薬企業の研究活動を支援することで、新日鉄ソリューションズ株式会社が貢献することを狙っている。

当社の得意分野である基盤設計構築技術とリレーショナル・データベースの活用技術をライフサイエンスの分野でどのように適用するべきか検討した。データを蓄積する層についての知見をためた後、ライフサイエンスに特有なデータへのアクセスのパターンについて知見を蓄積した。さらには、医学系大学との共同研究のなかで生活習慣病の原因遺伝子を解明する活動を支援した。これは取得したデータから統計的手法を用いてデータマイニングを行い、遺伝的要因とあわせて重要な重み付けを持つ、パラメータをいくつか選択するためのアルゴリズムを開発した。また、別の医学系大学との共同研究では質量分析計から出力される大量データを、解析しやすい形で蓄積するための基盤応用技術を現場適用した。

[研究成果の概要と今後の取り組み]

(3) リレーショナル・データベース技術の適用性検討

バイオインフォマティクス分野においては塩基配列などフラットファイルでデータを持つことが多い。ORACLE等のリレーショナル・データベースをこの分野に適応する際の設計・構築するべきかを検討した。検討結果として正規化が可能でリレーショナル・データーベースで検索速度の向上が期待できるデータ形式。正規化が困難で独特の検索アルゴリズムの適用が必要なものを区別して扱うことが必要であることがわかった。しかしながら、配列にＩＤを付与しそれらをもとに他の属性データと関連づけるなどの手法で配列データから関連する属性データを検索する(Blast検索)、属性データをもとに関連する配列データを検索するなど、リレーショナル・データベースと従来の検索手法を組み合わせることは有用であることが分かり、現状他の統合公共データベースでも主流になってきている。今後はリレーショナル・データーベースに配列データを格納した場合、テキスト検索機能へのBlast検索機能の融合した場合の有用性を検討したい。

(2) 基盤技術の応用

バイオインフォマティクスの周辺領域であるクリニカルインフォマティクス分野では、フラットファイル、RDB,XMLなどさまざまなデータの形式がある。これらのフォーマットを一元的に扱うためのミドルウェア利用方法を検討した。Oracleなどの商用データベース製品はフラットファイル、RDB、XMLなどを連携(Federate)して、検索する仕組みを提供している。これらのミドルウェアを利用するといままで個別にアクセスする仕組みを個別に開発していた工数を節減できることが期待できると考えていた。

ところが、現状分散するデータソースを分散したまま検索する技術はパフォーマンスの問題を多く抱えていることが分かり、実用段階にはまだ達していないという結論に至った。今後は64bit CPUが主流になり広いメモリー空間が活用できるようになると、対象データをオンメモリーで扱えるようになり、パフォーマンスの問題も改善できる可能性があると思われる。

(3) 遺伝子データ解析技術に関する研究

成人病の原因遺伝子と生活習慣因子との相関関係を分析するためにSNPsデータと多項目のクリニカルデータを対象にデータマイニングを行った。

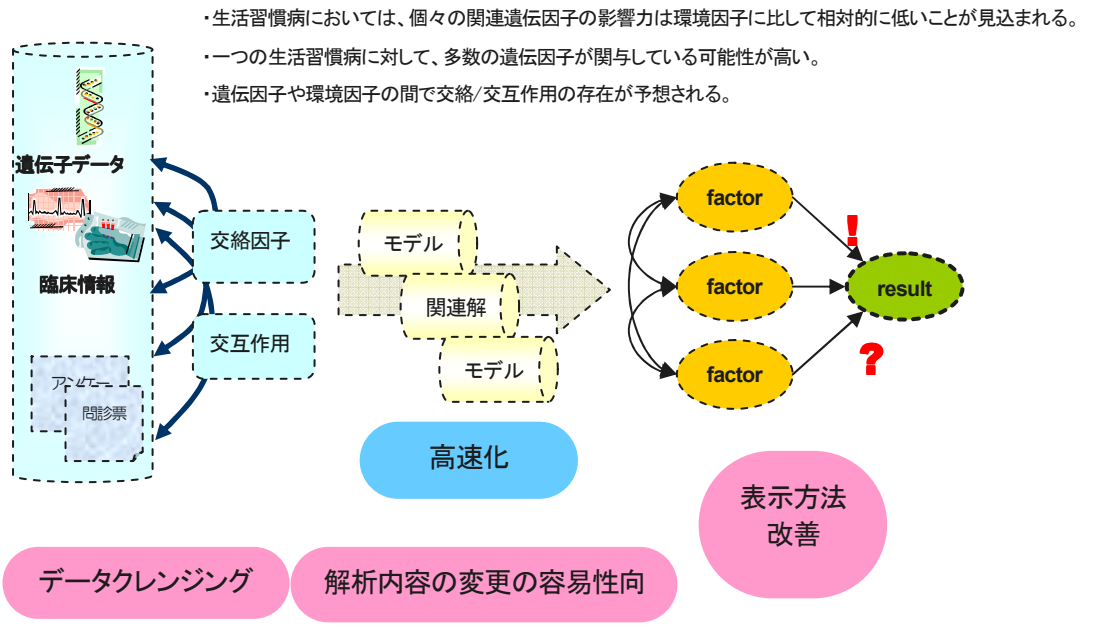

図 2-1 データマイニングによる要因分析の取り組み例

成人病は遺伝的要因のみならず複数の原因因子が関連して起きることが知られている。複数の原因因子を抽出するための独自のアルゴリズムを考案し、実データに適用した。今後は今回開発したアルゴリズムは遺伝子解析以外の因果分析に使用できるので、マーケティング分析などの他の分野のシステム構築案件に応用できないか、調査を検討していく予定である。

(4) 効率的なプロテオーム解析のための総合的研究環境を実現するシステム構築に関する研究/大規模データの分析・活用技術の研究タンパク質や遺伝子のデータなどの大量の蓄積データから有益な情報を効率よく抽出するためのシステム構築

臨床検体からとった病理サンプルをもとにタンパク質レベルでの病態プロテオミクス研究を実行する際に必要となるＩＴ基盤に関する調査研究を行った。

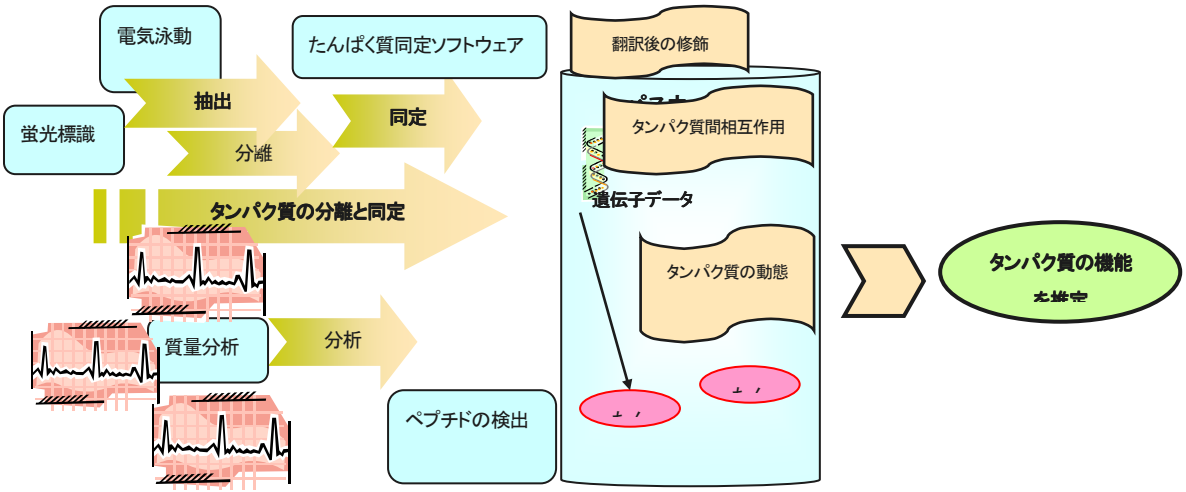

図2-2　プロテオーム研究におけるＩＴ基盤

複数の質量分析計から出るデータを効率よく管理するストレージ基盤を提供した。実験機器から出てくるフローデータをどうストックし再活用するかということは、急速に進歩しつつあるプロテオミクスの分野では新規性であった。

具体的には、NAS(Network Attached Storage)という仕組みを導入し、データが増加したときに、ユーザーが特に意識することなく拡張可能で、ディスクの二重化などが容易な仕組みを導入し、実験機器の周辺ＰＣに散在していたデータを取りまとめ、それまで研究活動ボトルネックであったデータ

のコピー作業を不要にすることができた。本研究は継続中であるので、今後はこの環境を前提としたアプリケーション基盤技術を検討していく予定である。