データ解析融合研究開発

医療を通じた社会への貢献に向けて

超高次元大規模ヘテロデータ解析技術と
生命体シミュレーションの融合

チームリーダー　宮野　悟（東京大学医科学研究所ヒトゲノム解析センター）
メンバー　秋山　泰（東工大情報理工学研究科計算工学専攻）
鎌谷直之（理研遺伝子多型研究センター）
樋口知之（統計数理研究所・副所長）

データ解析融合チーム

研究総括

東京大学（宮野　悟）

大規模遺伝子ネットワーク推定とその応用

大規模遺伝子ネットワーク推定技術	ベイズ的情報統合技術	生命体データ同化技術
連鎖・連鎖不平衡解析技術	ハプロタイプ解析技術	タンパク質ネットワーク予測技術

統計数理研究所（樋口知之）

生命体シミュレーションのためのデータ同化技術の開発

理化学研究所遺伝子多型研究センター（鎌谷直之）

大規模ゲノム多型データと表現型データを関連付ける新規アルゴリズムの開発と、妥当性、有用性の検討

東京工業大学（秋山　泰）

大規模タンパク質ネットワーク推定とその応用

データ解析融合チーム研究開発マップ

「一般」のデータ

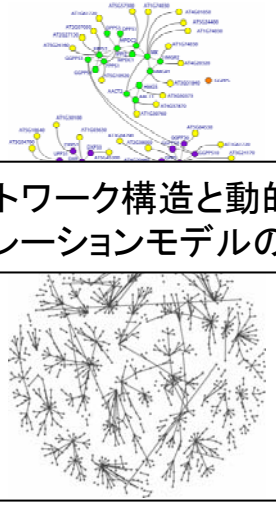
トランスクリプトーム、プロテオームデータをはじめとする生命システムの観測データ。

「個」のデータ

百～数万人の個人SNP及びゲノム配列。個人の疾患や薬物反応、質的形質データ。

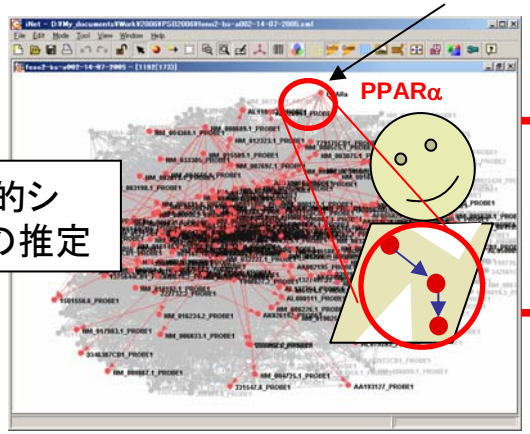
ペタフロップス級の計算によって、創薬ターゲット探索や個人差を考慮した医療開発に貢献する。

「一般」のモデル

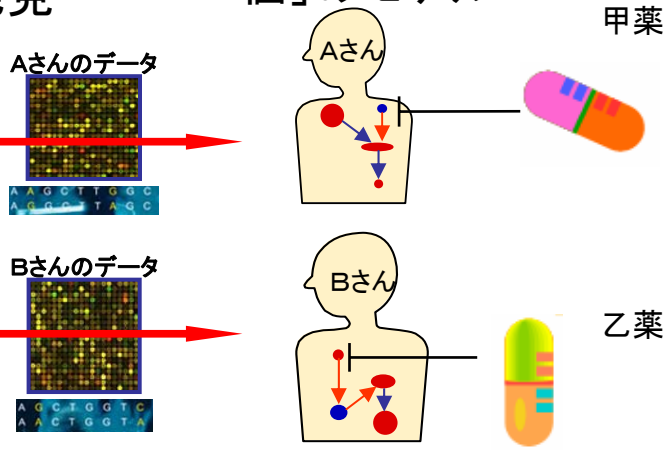


ネットワーク構造と動的シミュレーションモデルの推定

創薬ターゲットの発見



「個」のモデル



大規模な生命分子のネットワークを推定する技術を開発し、これを「地図」として薬物・疾患に関与する遺伝子群を探索。

「個」のデータを「一般」のモデルに合理的にフィットさせる生命システムのためのデータ同化技術の開発。

表現型（疾患や薬物反応性）に関連する遺伝子の解明と、個人の表現型をゲノム情報と環境情報により予測する技術。

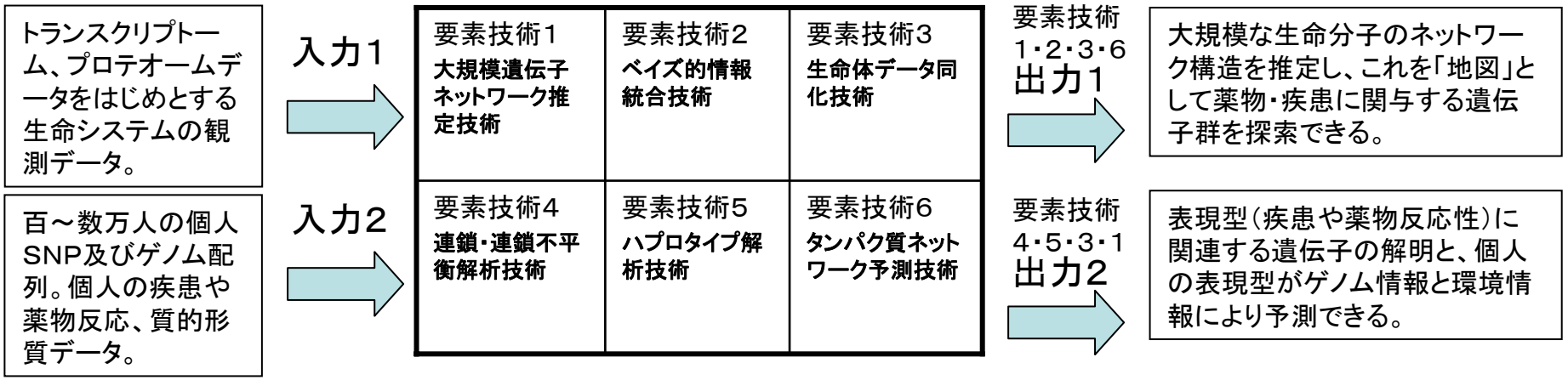
どんなソフトウェアが2012年までに開発されるのか

- ヒトの全遺伝子・転写産物を対象したネットワーク解析を可能にする大規模遺伝子ネットワーク推定ソフトウェア。
- PPIチャレンジ：1000×1000の超大規模計算を可能にする網羅的タンパク質間相互作用推定ソフトウェア（タンパク質ドッキング解析プログラム）。
- 50万SNPを自在に解析可能にする大規模SNP解析ソフトウェア。
- 「個」のデータを「一般」のモデルに合理的にフィットさせる生命システムのためのデータ同化を柱としたプログラム群。
- 以上を、融合し、統合的に利用活用するためのソフトウェア環境。

現状での計算規模と次世代スパコンでの計算規模

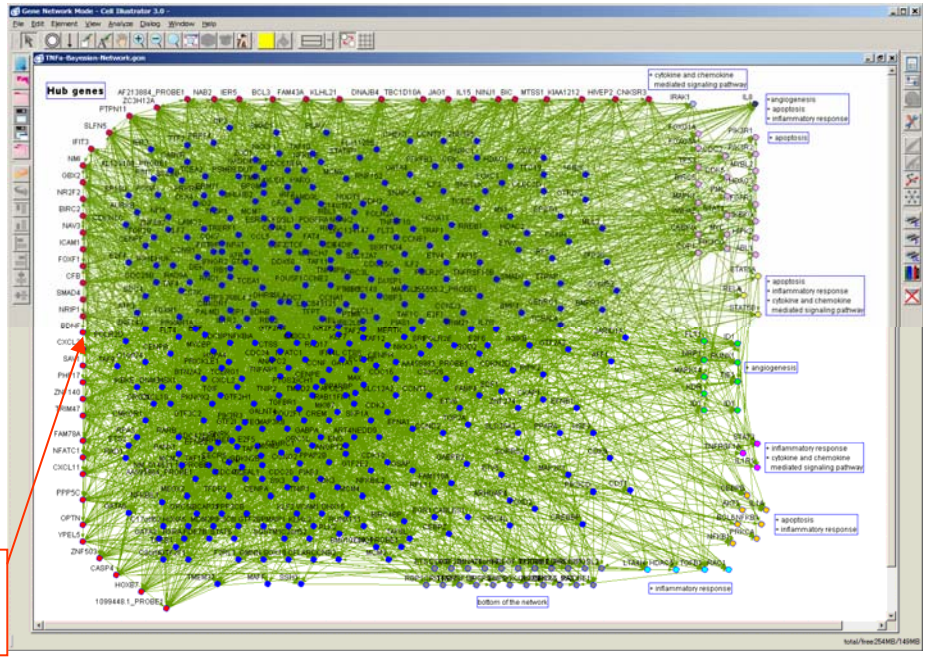
- Bayesianネットワークによる遺伝子ネットワークを利用した創薬ターゲット遺伝子のイン・シリコ探索では、128CPUのPCクラスタで、1000遺伝子（ヒト全遺伝子の3%程度）のネットワーク探索プロセスは、計算だけでも1ヶ月以上かかってきた。また大規模ネットワークが計算できても、創薬ターゲット探索のためにネットワークを解釈・解析するためのソフトウェア環境が未整備。数万から十数万の遺伝子転写産物を考えることは、現実的に不可能というのが現状。次世代スパコンでは、少なくともヒト全遺伝子を対象とした探索及び解析が可能になる。
- 創薬で興味の深いタンパク質を、（10万種の中から、その1%にあたる）1000種に絞ったとしても、1000×1000のタンパク質相互作用総当たりの評価では100万組の計算が必要であり、BlueGene(1/2ラック)で50年を要する。1PFLOPSの計算ができれば、2ヶ月で可能。現実では、10×10程度の計算にPCクラスタ（100CPU)で数時間が使われている。
- 連鎖解析では、6座位を用いて、広義の70,000表現型を取り扱い、ヒト全染色体を対象とすると、10TFLOPSのコンピュータで204日かかる。1PFLOPSの計算能力があると、連鎖解析は2.5日で解析でき、また9座位に広げることができる。ハプロタイプと表現型の関連解析では、座位数が20以上ある500個のハプロタイプブロックに対しては1,200日かかり、これが12日で解析可能となる。
- データ同化技術については、通常、生命の分子ネットワークシステムのデータポイント数が少ないため、十数個程度のパラメータに限定して様々な生物知識プライアを用いることで対応することが、現在のスパコンでの限界。次世代スパコンでは、この範囲を、ネットワーク構造の探索も含め数十パラメータに広げることができる。

どんなプログラムか


トランスクリプトーム、プロテオームデータをはじめとする生命システムの観測データ。
百～数万人の個人SNP及びゲノム配列。個人の疾患や薬物反応、質的形質データ。
入力1
入力2
要素技術1
大規模遺伝子ネットワーク推定技術
要素技術2
ベイズ的情報統合技術
要素技術3
生命体データ同化技術
要素技術4
連鎖・連鎖不平衡解析技術
要素技術5
ハプロタイプ解析技術
要素技術6
タンパク質ネットワーク予測技術
要素技術1・2・3・6
出力1
要素技術4・5・3・1
出力2
大規模な生命分子のネットワーク構造を推定し、これを「地図」として薬物・疾患に関与する遺伝子群を探索できる。
表現型（疾患や薬物反応性）に関連する遺伝子の解明と、個人の表現型がゲノム情報と環境情報により予測できる。

どんなプログラムか（1）

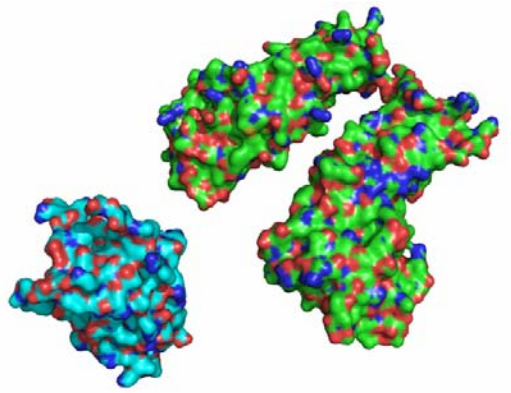
- **ヒトの全遺伝子・転写産物を対象したネットワーク解析を可能にする大規模遺伝子ネットワーク推定ソフトウェア**
  - 正則化法によるモデル推定
  - 大規模な生命分子のネットワークを推定・探索することが可能なプログラム
    - 入力：トランスクリプトーム、プロテオームデータをはじめとする生命システムの観測データ
    - 出力：大規模な生命分子のネットワーク構造とダイナミクス
- 何ができるか
  - 大規模な生命分子のネットワーク構造を「地図」として、シミュレーションと合わせて、薬物・疾患に関与する遺伝子群を探索できる
- 将来は
  - 創薬プロセスの基本ツールとして利用される。
- 誰が使うのか
  - 開発段階：ライフサイエンス関係の研究者及びベンチャー企業
  - 完成時には
    製薬関連企業及びシステムズバイオロジー研究者



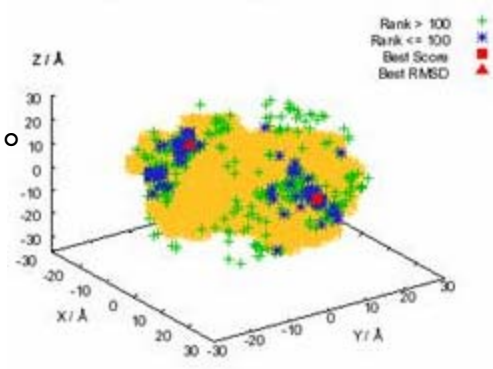
351種の遺伝子の siRNA ノックダウンによるDNAチップ解析データから推定されたヒト血管内皮細胞遺伝子ネットワークと同定されたハブ遺伝子群

どんなプログラムか（2）

- **タンパク質ドッキング解析プログラム**
  - 三次元物体の表面形状＆物理化学特性マッチング
    - 入力：専門形式（PDB)による2個のタンパク構造情報
    - 出力：ドッキング・ポーズの候補と評価スコア
  - 三次元複素配列間の畳み込み積分計算が中心
    - FFTの技法で $O(n^3 \log n)$の計算量で高速に実現する。



- 何ができるか
  - NMRやX線解析で解かれたタンパク質の構造から他のタンパク質とのドッキングの可能性を推定できる
  - 計算の高速化および超並列化により、既知の多くのタンパク質間での網羅的なドッキングの検査が可能に。
- 将来は
  - NMRやX線解析で立体構造が得られた直後に、既知のタンパク質とのドッキングの可能性が示唆され、実験者に確認実験を示唆できる。また、計算機による立体構造予測の結果についても、ドッキングの可能性のチェックができる。
- 誰が使うのか
  - 開発段階
    タンパク質構造研究者
  - 完成時には
    製薬会社による創薬ターゲット探索



どんなプログラムか（3）

- 50万SNPを自在に解析可能にする大規模SNP解析ソフトウェア
  - 多座位のゲノム多型データを用いた高速計算のための 連鎖解析プログラム
  - 全ゲノムをカバーするSNP遺伝子型を用いた高速計算のための 関連解析プログラム
    - 入力：百～数万人の個人SNP及びゲノム配列。個人の疾患や薬物反応、質的形質データ、及びトランスクリプトーム、プロ テオームデータをはじめとするシステムの観測データ
    - 出力：→なにができるのか
- 何ができるのか・将来は
  - 表現型（疾患や薬物反応性）に関連する遺伝子の解明と、個人の表現型がゲノム情報と環境情報により予測
- 誰が使うのか
  - 開発段階：先端医療開発の現場
  - 完成時には：先端医療開発の現場及び医療の現場

どんなプログラムか（４）＆（５）

- **生命体データ同化プログラム**
  - 「個」のデータを「一般」のモデルに合理的にフィットさせる生命システムのためのデータ同化を柱としたプログラム群
- 何ができるのか・将来は・誰が使うのか
  - セッション①樋口知之：データ解析融合「階層を越えるモデリングへの挑戦」を参照

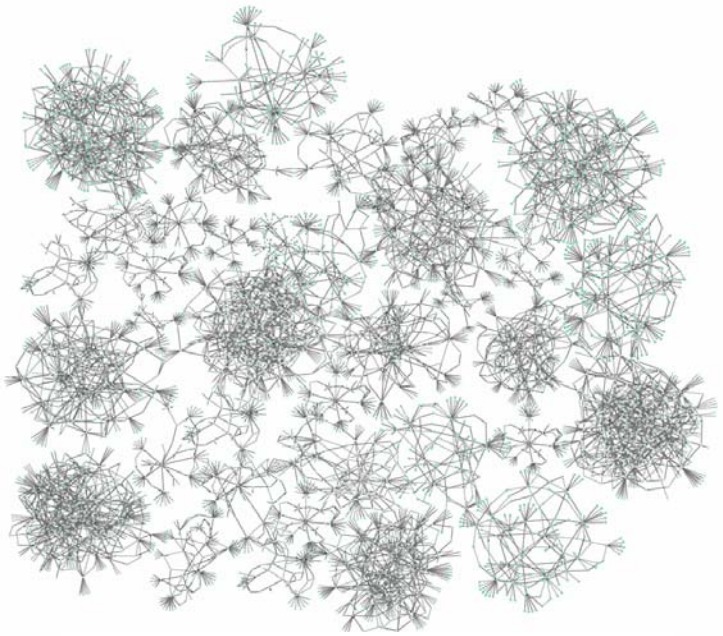
- **融合・統合ソフトウェア環境**
  - 既存の商用・非商用GUIソフトウェア及びデータベースと本研究で開発されたソフトウェアを統合したソフトウェア環境の構築
    - Cell Illustrator、BIOBASEなどを利用
- 何ができるのか・将来は・誰が使うのか
  - ペタフロップス級の計算によって、創薬ターゲット探索や個人差を考慮した医療開発の支援ソフトウェアプラットフォームとして機能

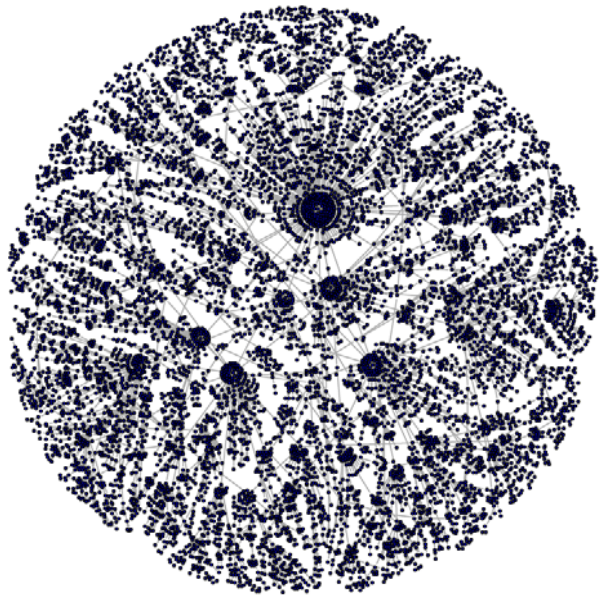
これまでの開発状況

大規模遺伝子ネットワーク推定とその応用

大規模遺伝子ネットワークのグローバル解析プログラムの開発

- L1正則化法であるWeighted Lassoを用いて、グラフィカルガウシアンモデルを推定するプログラムを開発。これにより、大規模遺伝子ネットワーク探索の最初のフィルターができた。
- ベイジアンネットワークを含む一般的な大規模遺伝子ネットワーク探索のアルゴリズムを開発した。
- ベイジアンネットワークや状態空間モデルなどに基づく大規模遺伝子ネットワークの精緻化戦略の構築が可能となった。



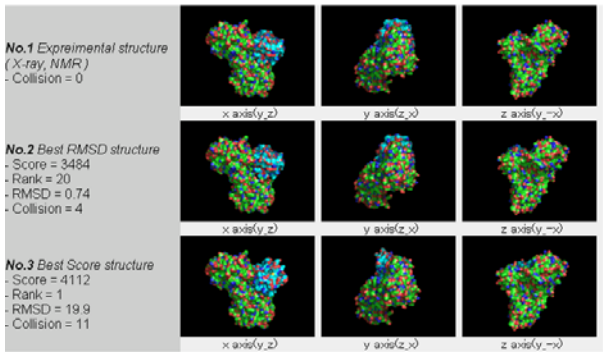


10,000ノードのネットワークから小規模ネットワークへのフォーカスが可能遺伝子間の関わりを大規模に解析できる。シミュレーションデータにより、この方式の高い精度が確認された。

大規模タンパク質ネットワーク推定とその応用

タンパク質ドッキング解析プログラムの開発

- エンジン部として、新たな三次元複素FFTプログラムを開発中。後処理部のためのクラスタリングによる解候補の選別方式の開発、および評価関数として経験的残基間ポテンシャルの新規採用。



- 網羅的実行環境構築、テスト集合準備
  権利関係に配慮したプログラム再設計
  - 新たな三次元複素FFTプログラム作成
  - 新たな後処理部の作成、TSUBAMEシステム移行

- H20年以降の、大規模実施による部分的相互作用ネットワーク解析及びポテンシャル改良による実効的精度保証へ向けた準備

大規模ゲノム多型と表現型データを関連付ける新規アルゴリズムの開発と、妥協性、有用性の検討

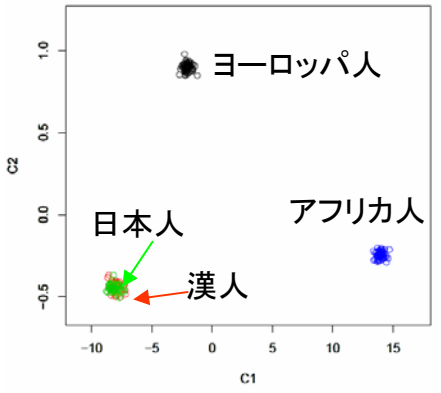
やはり10－50万SNPの威力はすごい。

- 連鎖解析アルゴリズムの高速計算システムでの実行
  - 2-3塩基の繰り返しであるマイクロサテライトを用いてメンデル型遺伝病（パラメトリック連鎖解析）、多因子形質（ノンパラメトリック連鎖解析）の座位を解明するアルゴリズムを高速計算に適したように改良（10^{100}のsample space）
- SNPハプロタイプを用いた連鎖解析アルゴリズムの高速計算システムでの実行
  - SNPハプロタイプを用いてメンデル型遺伝病（パラメトリック連鎖解析）、多因子形質（ノンパラメトリック連鎖解析）の座位を解明するアルゴリズムを高速計算に適したように改良

⇒これにより全ゲノムから単因子遺伝要因の解明できる。

- 種々の多段階関連解析の手法の高速化
  - Replication法、Joint法、P値積法
- Permutation法を用いた関連解析の高速化（10^{600}のサンプル空間）
  - 表現型Permutationにより帰無仮説での統計量の分布を計算
- 正確計算法を用いた関連解析の高速化（10^{100}のサンプル空間）
  - 帰無仮説での正確な確率を計算
- 関連解析を用いた遺伝子間相互作用解析の高速化
  - 2つ以上の多型の表現型への影響の相互作用を解析
- 個人の薬物反応性を正確に予測する高速計算のための計算技術を今後開発

16万SNPを使ったMDSを用いたヒトのクラスタリング解析



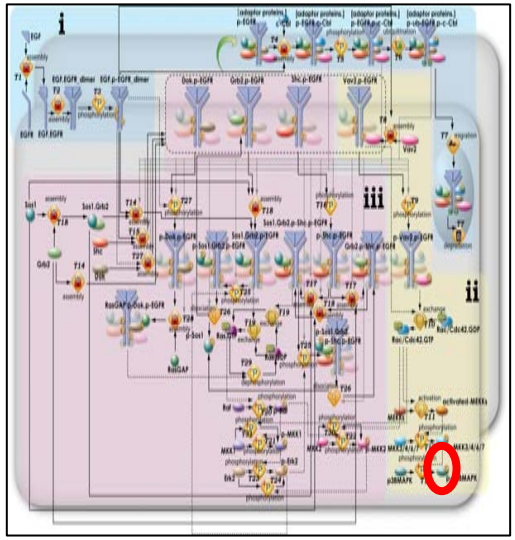
染色体1-22のSNP を用いたMDSによるクラスタリング
SNP 数 9212
Sample 数 210
処理時間 10 分程度

これまでの開発状況

生命体シミュレーションのためのデータ同化技術の開発

生命体データ同化技術によるシミュレーションモデルとデータの融合が小規模で可能に。

癌などの病気に関わっているEGF受容体を介したシグナル伝達系のシミュレーションモデルの構築(Cell Illustratorを使用）



63パラメータのモデル

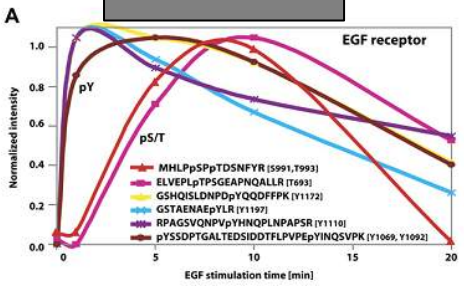
モデル

SILAC法による量的プロテオーム時系列データをモデルに同化

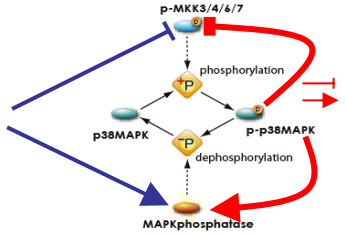
データ同化技術

●現在のスパコンでは20パラメータが限界。3つに分解して対応。
●ペタコンで数十パラメータを達成。

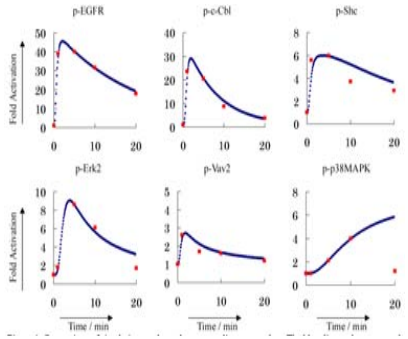
データ



新たな機序仮説の生成

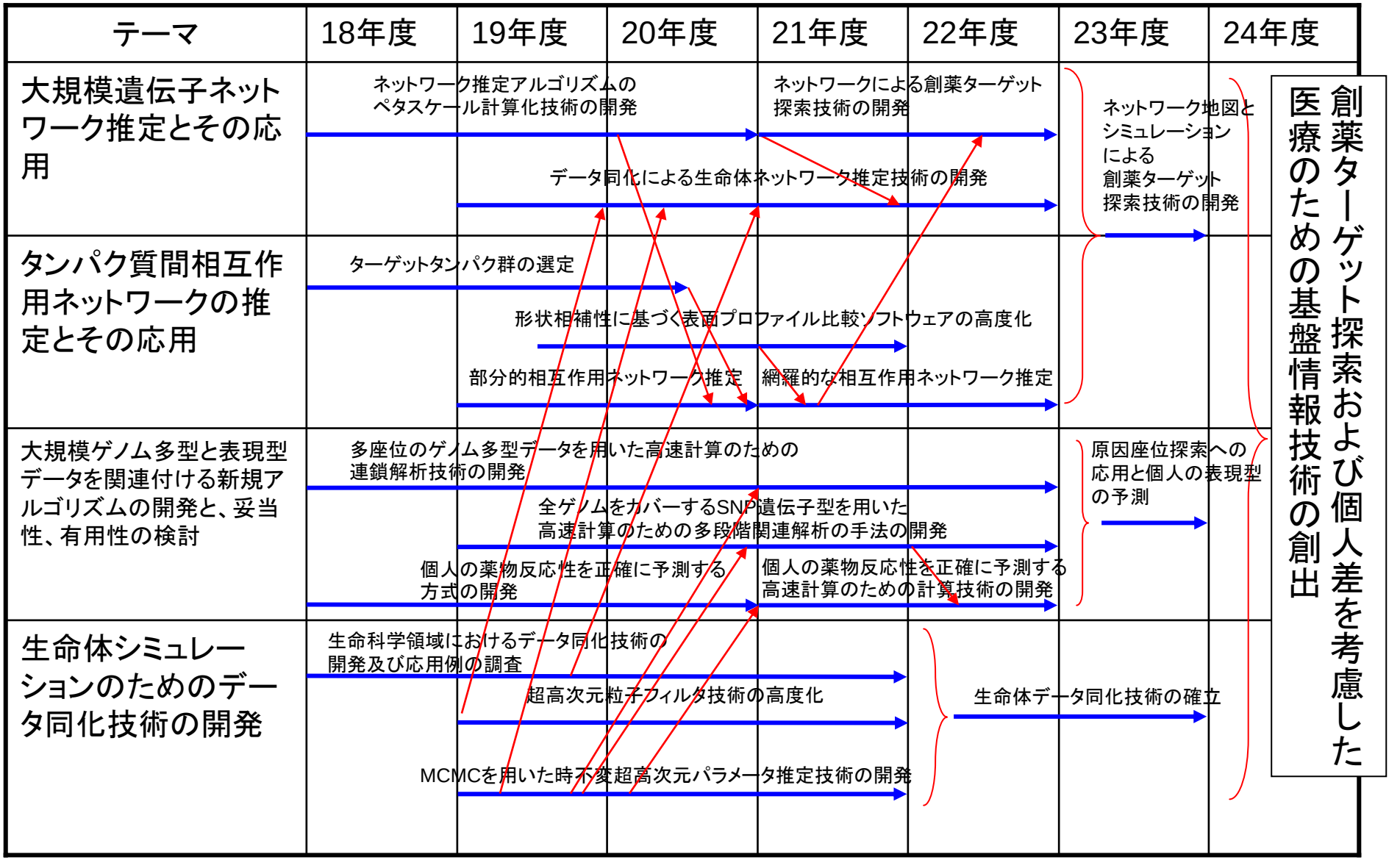


Hypothesized regulations



予測とエラーの発見

これまでの開発状況と今後の計画



2006年～2010年に達成すること	2011年～2012年度に達成すること
●ネットワーク地図とシミュレーションによる創薬ターゲット探索技術の開発 ●生命体データ同化技術の確立 ●原因座位探索への応用と個人の表現型の予測方式の開発 　これら三の目標を達成するための要素技術開発とプログラム開発を行う。	2010年度までに開発した技術とプログラムを用いて、次の目標を達成する。 ●創薬ターゲット探索および個人差を考慮した医療のための基盤情報技術の創出

将来は

- ペタスケール計算により、遺伝子情報のデータ解析および生命システムモデリングにおいて、明らかに、別世界が出現する。
- 特に、データ同化については、生命科学・医学研究においてはじめてのチャレンジであり、ペタスケール計算があってはじめて実現できるもので、生命体シミュレーションに大きなインパクトを与える可能性がある。
- 大規模な生命分子のネットワークを推定・解析することが可能になり、これを「地図」として薬物・疾患に関与する遺伝子群の探索が可能になる。この地図上で、化合物の影響パスウェイを探索・シミュレートしたり、NMRやX線解析で得られたタンパク質の立体構造情報とあわせて、それらのネットワーク上およびその他のタンパク質とのインタラクションをイン・シリコで予測することができる。
- そして、創薬プロセスにパラダイムシフトを起こすとともに、大規模SNP解析とシミュレーションによる「個」にフィットした医療開発に貢献できる。