データ解析融合研究開発チーム成果報告

データ解析融合研究開発チーム

宮野　悟

次世代計算科学研究開発プログラム
データ解析融合研究開発チーム チームリーダー

発表者紹介

1977 年 3 月　九州大学理学部数学科卒業
1979 年 3 月　九州大学大学院理学研究科修士課程数学専攻修了
1979 年 6 月　九州大学理学部助手
1985 年 4 月　Alexander von Humboldt 財団研究員
1987 年 4 月　Universität GH Paderborn 助手
1987 年 12 月　九州大学理学部助教授
1993 年 3 月　九州大学理学部教授
1996 年 4 月　東京大学医科学研究所教授

研究分野

計算システム生物学、バイオインフォマティクス、メディカルインフォマティクス

データ解析融合研究開発チーム

宮野　悟
次世代計算科学研究開発プログラム
データ解析融合研究開発チーム チームリーダー

1. 目的

計測技術の大規模化・精緻化・簡便化により、遺伝子から環境因子にわたる生命システムに関するデータが、超高次元化・超ヘテロ化・超膨大化した。しかも、不観測性・欠損の問題も同時に有している。こうした中、データ解析はデータの多様化と増大化に急速に引き離され、シミュレーションは生命体個別の現実データを反映できず予測能力に限界がある。方法論にパラダイムシフトが必要となった。

そこで本研究チームは、ペタスケールの計算を前提にして、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を融合し、生命体システムに対する予測と発見の基盤情報技術を構築することを目的として研究開発を行っている。４人の PI が、図１の文脈で、次の３つの技術開発を目標に研究を行ってきた。

① ゲノムワイド関連解析により疾患や薬物反応性などの表現型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
- 理化学研究所ゲノム医科学研究センター・角田達彦(2011.4～)；鎌谷直之(2006.10～2011.3)

② 大規模な生体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する技術
- 東大医科学研究所・宮野悟
- 東工大・秋山泰

③ データ同化技術の活用による一般のモデルから個のモデルを創出する技術
- 統計数理研究所・樋口知之

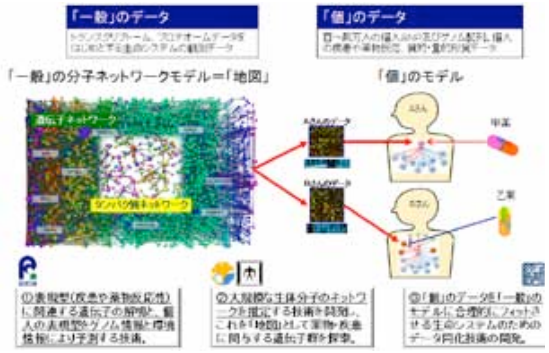


図 1

2. 現時点での成果

2.1 「肺がんと薬」を研究の共通軸する

本チームの研究開発は 2006 年に開始したが、2008 年に実施された研究の進捗状況についての中間評価において、チームとして共通の研究軸をもつことによりチーム内での相乗効果を発揮させることが指示された。そこで、本チームのメンバーが実験データの取得も含め対応可能なものとして「肺がんと薬」を共通のテーマとした。そして、「肺がんと薬」をシステムを理解するための課題として、図２のように技術開発の方向を明確化した。

2.2 チームの構成と役割・開発ソフトウェア

図３のチームの構成と役割により、おもに５つに分類されるソフトウェア SiGN、MEGADOCK、ParaHaplo、LiSDAS、SBiP を開発しており、京コンピュータ上での現時点で許された最大高並

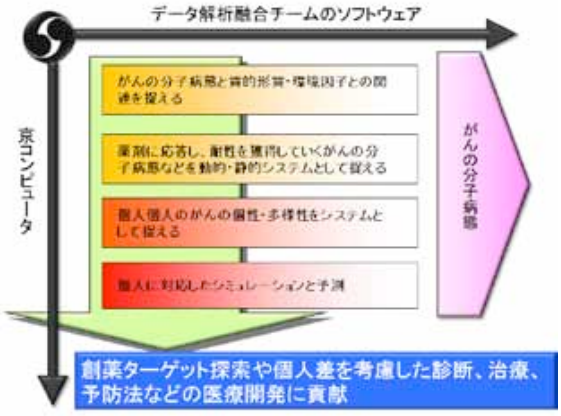


図 2

列化（12280 ノード)を一部達成している。

SiGN は、ヒト全遺伝子規模の遺伝子ネットワークを推定・探索するソフトウェアで、ベイジアンネットワーク、状態空間モデル、L1 正則化法を駆使したネットワーク推定法に基づいたアプリケーション群がパッケージに含まれている。トランスクリプトームのネットワークを推定することに用いることができるが、ソフトウェアとしては生命科学に限らず、汎用的なものである。MEGDOCK は、1000×1000 規模のタンパク質相互作用を網羅的に推定することでタンパク質ネットワークを構築できるソフトウェアである。網羅的なタンパク質間相互作用予測システムを京コンピュータの上で効率的に動作する並列ソフトウェアパッケージとしてまとめ、前処理・後処理のツール群も京コンピュータ周辺環境で動作するよう移植を行っている。MEGADOCK で予測されたタンパク質相互作用情報をトランスクリプトームネットワークと合わせてシステムを理解することを目指している。これらのソフトウェアを使ったデータ解析により、肺がんなどの分子ネットワーク・薬剤応答ネットワークを解析し、その分子病態を描出することを狙っている。他の疾患や薬にも応用可能である。

ParaHaplo は、肺がんに限ったものではなく、患者群と対照群の全ゲノム上の SNP を用いハプロタイプ単位のゲノムワイド関連解析を行い、疾患関連遺伝子を網羅的に探索するためのソフトウェアである。ファーストランナーとして京コンピュータ上で並列度を上げている。京コンピュータ上で実行させて、実際に肺がんや、その治療薬などに関係する形質を含め、新たな疾患関連遺伝子、形質関連遺伝子を発見する成果を目指している。これに関連して、拡張 RAT 法による 2 SNP 組合せの全ゲノム関連解析ソフトウェア（ExRAT）を開発している。これは、病気へのリスクを上げる原因となるような複数の遺伝子による相乗効果を大規模体系的網羅的に見つけ出すソフトウェアであり、京コンピュータのアーキテクチャに合わせ並列度を上げ、京コンピュータ上で実データを用いて実行させて、未知の複数要因による新たな疾患発症機序の発見と、未だ解決してない遺伝力の問題を解く成果を狙っている。また、次世代シークエンサーデータ解析プログラム（NGSanalyzer）は、がんを体細胞変異のゲノム学と遺伝統計学により解明するために、次世代シークエンサーからの人間のがん細胞と正常細胞の全ゲノムの膨大なデータを網羅的・正確・高速に解析するためのソフトウェアである。現在、

図 3

パイプラインの並列度を上げ、京コンピュータ上で実行させて、実際のがんのメカニズムと日本人のゲノムの多様性の解明を目指している。こうした解析から見つかる遺伝子を、SiGN や MEGADOCK で推定したネットワーク（地図）にマップし、関連解析の結果を分子ネットワークで理解することを狙っている。

データ同化技術は、個別データをモデルに同化させ、個別化モデルを作り、個に対するシミュレーションと予測をするものである。LiSDAS は、階層性を考慮した粒子フィルタアルゴリズムなどを京コンピュータ上で超高並列化したソフトウェアである。オミックス計測技術から生成されるデータを利用して、生体内分子相互作用ネットワークのシミュレーションモデルを構築することに応用することを目指している。

開発したソフトウェアを連携して使うことができる SBiP (Systems Biology integrative Pipeline)というデータ解析プラットフォームを開発している。SBiP は京コンピュータで走らせるソフトウェアではなく、ユーザのコンピュータにインストールし、高機能 GUI により、SBiP から京コンピュータの上で走るソフトウェアをつないで解析パイプラインを作成できるものである。

また、おもに SiGN と LiSDAS の開発を進めるために肺がんと抗がん剤ゲフィチニブに関する mRNA などの時系列データを若干取得し、活用している。

3. プロジェクト終了時の達成目標

ネットワーク解析による薬のターゲット探索やがんなどの病態を理解する情報技術、並びにゲノム情報に基づいた個人差を考慮した医療のための基盤情報技術を創出する。

データ解析融合研究開発チーム
チーム成果統括報告

理化学研究所
次世代計算科学研究開発プログラム
データ解析融合研究開発チーム　チームリーダー
宮野　悟


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


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（角田達彦）
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

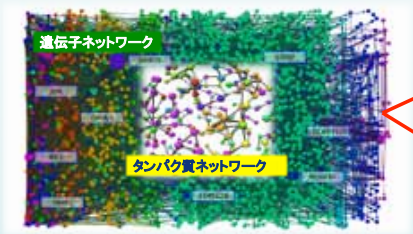

東京工業大学（秋山　泰）
大規模タンパク質ネットワーク推定とその応用

データ解析融合チームの背景・目的

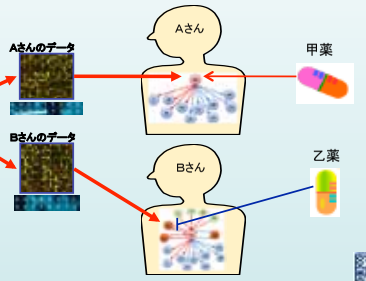
「一般」のデータ
トランスクリプトーム、プロテオームデータをはじめとする生命システムの観測データ

「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・量的形質データ

「一般」の分子ネットワークモデル＝「地図」



「個」のモデル



①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。

②大規模な生体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。

③「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。

「肺がんと薬」のシステムを理解するための課題

データ解析融合チームのソフトウェア

京コンピュータ

- がんの分子病態と質的形質・環境因子との関連を捉える
- 薬剤に応答し、耐性を獲得していくがんの分子病態などを動的・静的システムとして捉える
- 個人個人のがんの個性・多様性をシステムとして捉える
- 個人に対応したシミュレーションと予測

→ がんの分子病態

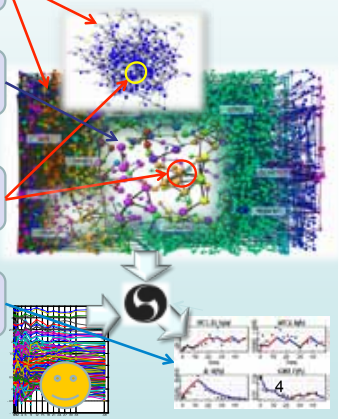
創薬ターゲット探索や個人差を考慮した診断、治療、予防法などの医療開発に貢献

概要・アプローチ
チームの構成と役割・開発ソフトウェア

SBiP Systems Biology Integrative Pipeline

大規模遺伝子ネットワーク推定（SiGN）宮野悟（東大医科研）	遺伝子ネットワーク	ヒト全遺伝子規模にわたる遺伝子ネットワークの推定・探索
タンパク質ドッキング解析（MEGADOCK）秋山　泰（東工大）	タンパク質ネットワーク	1000x1000タンパク質相互作用推定でネットワークを網羅的に解析
大規模SNP関連解析（ParaHaplo, ExRAT, NGSanalyzer）角田達彦（理研）	ゲノムバリエーション	肺がんの特性、抗がん剤の効果・副作用と遺伝子との関連付け
生命体データ同化（LiSDAS）樋口知之（統数研）	シミュレーション	個別データをモデルに同化させ、個別化モデルを作る

肺がんの分子ネットワーク・薬剤応答ネットワークを解析し肺がんの分子病態を描出



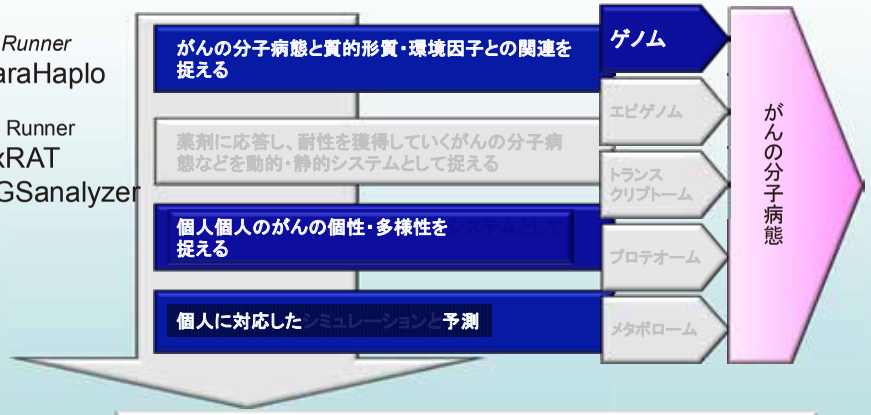
「肺がんと薬」データの取得

開発アプリケーション

アプリケーション名	略称	担当PI
ハプロタイプ関連解析に於ける統計検定を行うためのソフトウェア	ParaHaplo	角田達彦
次世代シークエンサーデータ解析プログラム	NGS analyzer	角田達彦
拡張RAT法による2SNP組合せの全ゲノム関連解析ソフトウェア	ExRAT	角田達彦
大規模遺伝子制御ネットワーク推定プログラム	SiGN SiGN-BN, SiGN-SSM, SiGN-L1	宮野　悟
データ解析融合プラットフォーム	SBiP	宮野　悟 スパコンには実装しない
生命体データ同化プログラム	LiSDAS	樋口知之
網羅的タンパク質ドッキング解析プログラム	MEGADOCK	秋山　泰

5

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　　担当　角田達彦（理研）



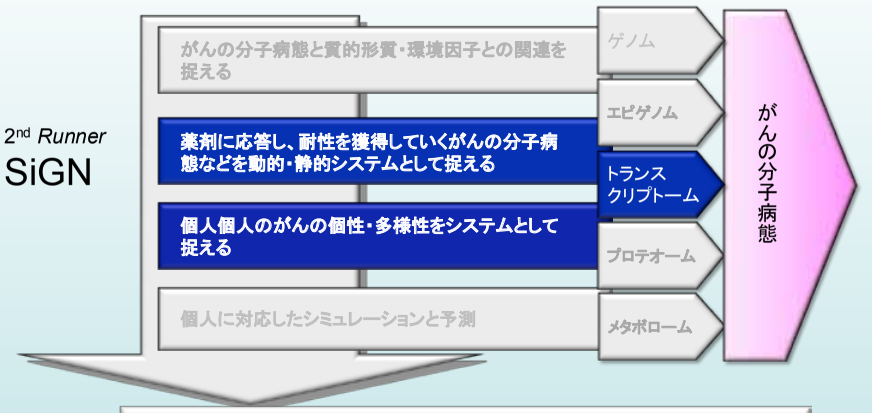
6

1st *Runner*　ParaHaplo
2nd *Runner*　ExRAT　NGSanalyzer

- **ParaHaplo**　肺がんに限ったものではなく、患者群と対照群の全ゲノム上のSNPを用いハプロタイプ単位のゲノムワイド関連解析を行い、疾患関連遺伝子を網羅的に探索するためのソフトウェア。肺がんや、その治療薬などに関係する形質を含め、新たな疾患関連遺伝子、形質関連遺伝子を発見する成果を目指している。
- **ExRAT**　拡張RAT法による2SNP組合せの全ゲノム関連解析ソフトウェア。病気へのリスクを上げる原因となるような複数の遺伝子による相乗効果を大規模体系的網羅的に見つけ出す。
- **NGSanalyzer**　次世代シークエンサーデータ解析プログラム。がんを体細胞変異のゲノム学と遺伝統計学により解明するために、次世代シークエンサーからの人間のがん細胞と正常細胞の全ゲノムのデータを網羅的・正確・高速に解析するためのソフトウェア。がんのメカニズムと日本人のゲノムの多様性の解明を目指している。

7

大規模遺伝子ネットワーク推定とその応用
担当　宮野　悟（東大）



8

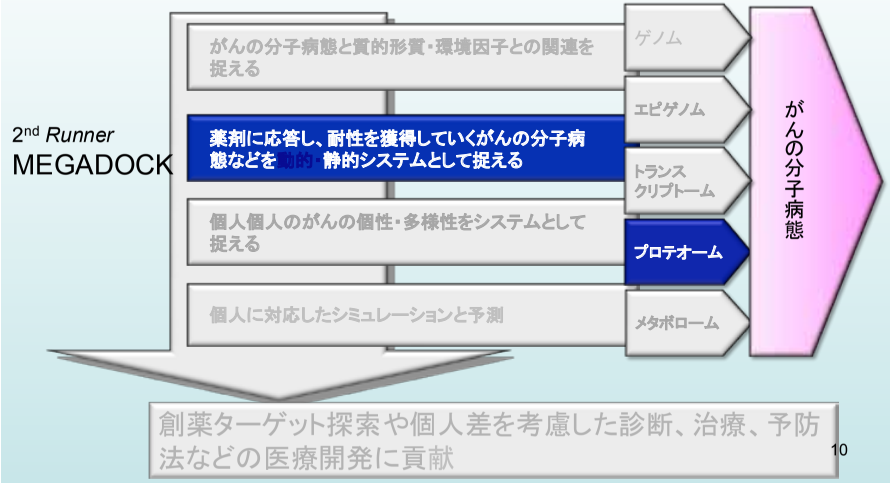
2nd *Runner* SiGN

SiGN-BN, SiGN-SSM, SiGN-L1

- ヒト全遺伝子規模の遺伝子ネットワークを推定・探索するソフトウェアで、ベイジアンネットワーク(SiGN-BN)、状態空間モデル(SiGN-SSM)、L1正則化法を駆使したネットワーク推定法(SiGN-L1)に基づいたアプリケーション群がパッケージに含まれている。
- トランスクリプトームのネットワークを推定することに用いることができるが、ソフトウェアとしては生命科学に限らず、汎用的なものである。

9

大規模タンパク質ネットワーク推定とその応用　担当　秋山　泰（東工大）



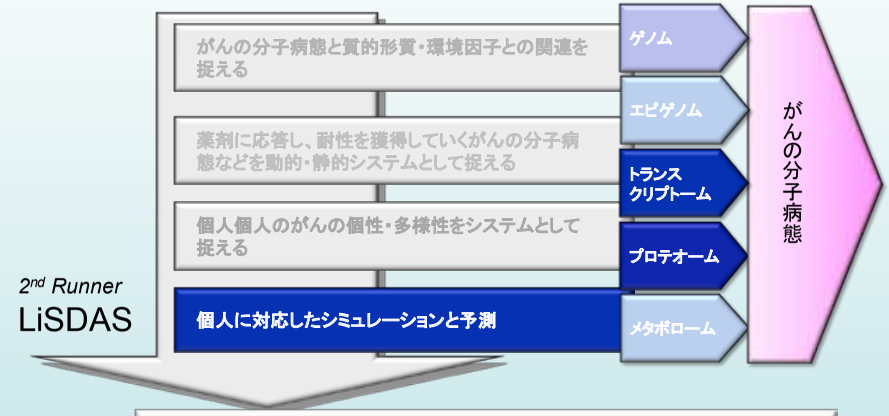
10

2nd *Runner* MEGADOCK

- 1000×1000規模のタンパク質相互作用を網羅的に推定することでタンパク質ネットワークを構築できるソフトウェア。MEGADOCKで予測されたタンパク質相互作用情報をトランスクリプトームネットワークと合わせてシステムを理解することを目指している。
- これらのソフトウェアを使ったデータ解析により、肺がんなどの分子ネットワーク・薬剤応答ネットワークを解析し、その分子病態を描出することを狙っている。

11

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　樋口知之（統数研）

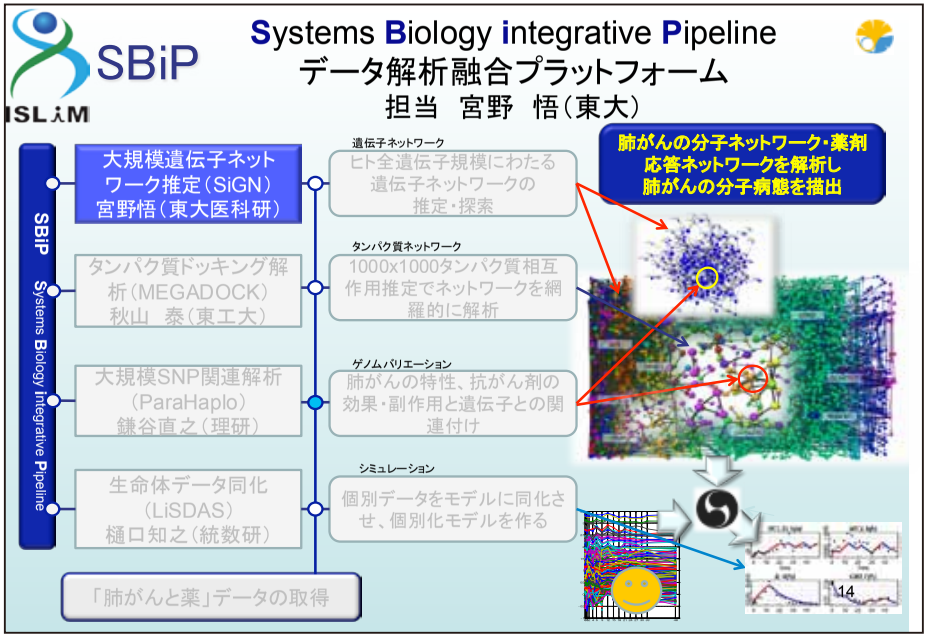


12

2nd *Runner* LiSDAS

- データ同化技術は、個別データをモデルに同化させ、個別化モデルを作り、個に対するシミュレーションと予測をするものである。LiSDASは、階層性を考慮した粒子フィルタアルゴリズムなどを京コンピュータ上で超高並列化したソフトウェアである。
- オミックス計測技術から生成されるデータを利用して、生体内分子相互作用ネットワークのシミュレーションモデルを構築することに応用することを目指している。

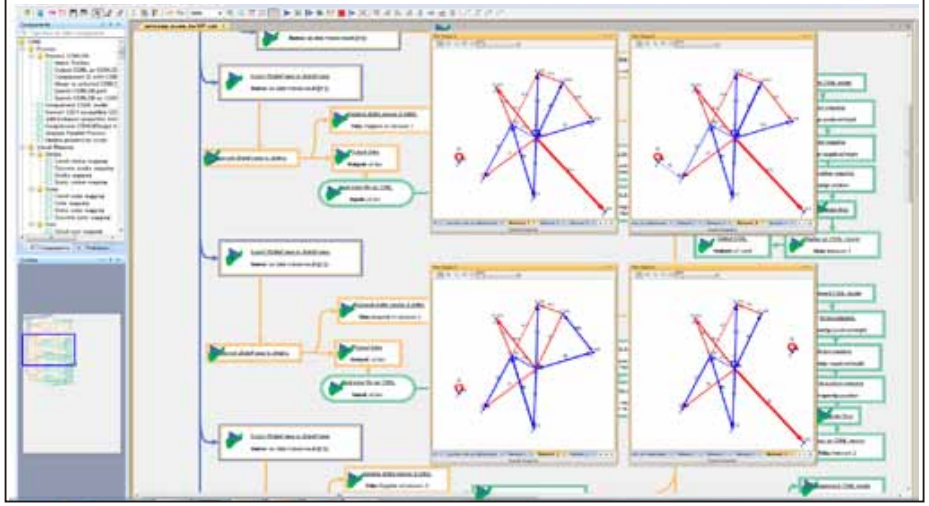
13

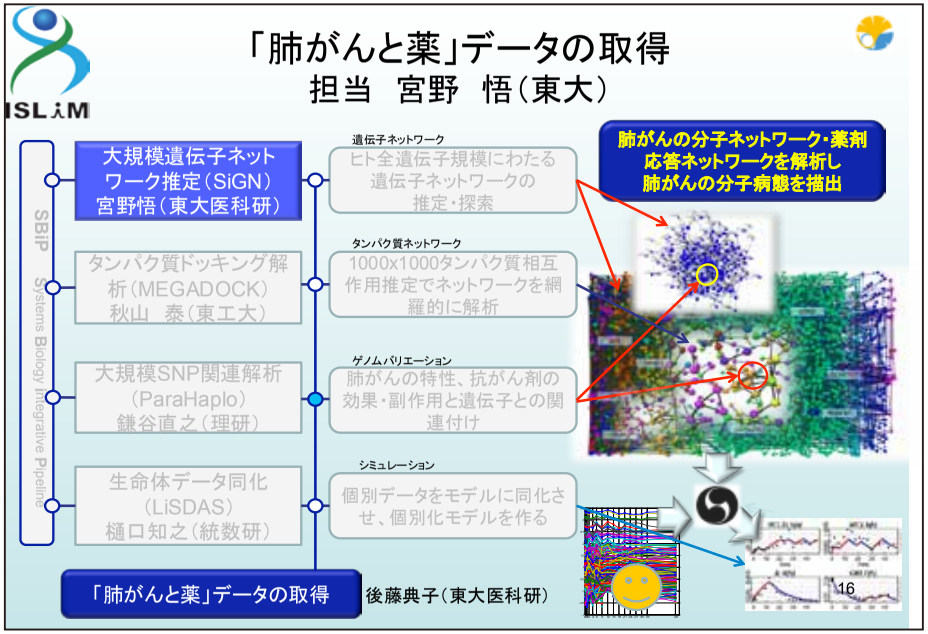


SBiPのスナップショット

ネットワーク推定と比較:

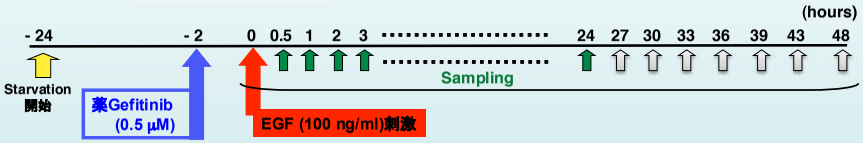
SBiPは京コンピュータで走らせるソフトウェアではなく、ユーザのコンピュータにインストールし、高機能GUIにより、SBiPから京コンピュータの上で走るソフトウェアをつないで解析パイプラインを作成するツール





「肺がんと薬」に関するデータの取得

細胞
■正常ヒト小気道上皮細胞
SAEC52, SAEC97
■肺腺がん細胞株PC9
■Gefitinib耐性株PC9GRM2、他



4つの実験条件

mRNA
タンパク質

17

現時点での達成状況と課題

【進捗状況】

•ソフトウェアに関しては、独創的な手法が開発され、優先順位をつけて京での高並列化がほぼ順調に進んでいる。現時点で可能な最大規模に達しているものもある。

【課題】

•4PIの成果をつないで初めて得られるようなインパクトのある成果がまだ出せていない。

プロジェクト終了時の達成目標

- 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
- ヒトの全遺伝子・転写産物を対象と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
- 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。
- 「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

- ネットワーク解析による薬のターゲット探索やがんなどの病態を理解する情報技術
- ゲノム情報に基づいた個人差を考慮した医療のための基盤情報技術

19

全ゲノム解析の超並列化による疾患研究の加速

角田 達彦
理化学研究所 ゲノム医科学研究センター
統計解析・技術開発グループ グループディレクター

発表者紹介

1995 年 3 月　東京大学大学院 工学系研究科 情報工学専攻博士課程修了
1995 年 4 月　京都大学大学院 工学研究科 助手
1997 年 4 月　東京大学医科学研究所 ヒトゲノム解析センター リサーチアソシエイト
1998 年 4 月　東京大学医科学研究所 ヒトゲノム解析センター 助手
2000 年 4 月　理化学研究所 遺伝子多型研究センター チームリーダー
2008 年 4 月　理化学研究所 ゲノム医科学研究センター チームリーダー
2011 年 4 月　理化学研究所 ゲノム医科学研究センター グループディレクター

医学博士・工学博士

研究分野

ゲノム医学，遺伝統計学，メディカルインフォマティクス

全ゲノム解析の超並列化による疾患研究の加速

理化学研究所
ゲノム医科学研究センター
グループディレクター
角田 達彦

背景・目的

- 背景: 全くの未知のものも含め，疾患の原因を探索するには，全ゲノム上で患者群と対照群との間でゲノムDNA配列を比較する，ゲノムワイド関連解析（GWAS）が極めて有効であり，われわれ理化学研究所ゲノム医科学研究センターでは，2002年に世界に先駆け初めてのGWASを実現してから，世界を牽引してきた．この方法論をさらに劇的に推進するために，単点の解析だけでなく，近傍の，あるいは遠距離にある複数点の解析を行ったり（それぞれParaHaploとExRAT），対象とするマーカーを，これまでの固定されたセットから拡張し，次世代シークエンサーにより全ゲノム配列への探索を行ったり（NGS analyzer）するなどの新たな方法論を導入する必要がある．これらには，スーパーコンピューティングを駆使することが必要不可欠になる．
- 目的:
  - ParaHaplo: GWASの方法論を拡張し，単点ではなく連続する複数点に対して超並列に行うことによって，飛躍的な検出力の向上と高速性能を狙う．
  - ExRAT: 複数の遺伝子の相互作用により発症する可能性を，全ゲノム上で網羅的に超並列に探索する．
  - NGS analyzer: 次世代シークエンサーの爆発的データを超並列に処理し，ゲノム多様性を高精度かつ包括的に解析する．

概要・アプローチ

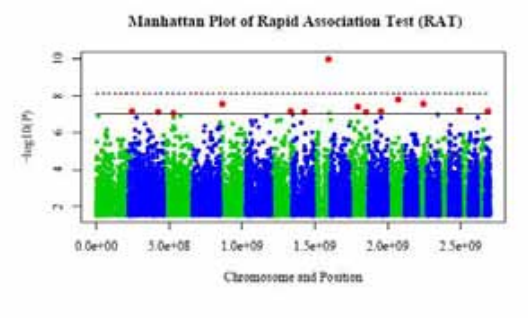
- 研究開発コードの概要
  - ParaHaplo: 人のゲノム全体にわたる遺伝的な相違点を，近傍にある複数点を同時にみたハプロタイプという単位で，患者群と対照群とで比較することにより，疾患の遺伝的原因を探るための統計検定計算
  - ExRAT: 遺伝子間相互作用が発症リスクを変化させる疾患関連遺伝子の組合せを全ゲノムで探索する．2SNP間の全組合せを超並列に行う方法と，SNP間の連鎖不平衡（相関）も考慮した，より精密な方法の2種類を実装．前者で全組合せをスクリーニングし，後者で経験的p値を求める手順を想定している．
  - NGS analyzer: 次世代シークエンサーの出力データを高速に解析し，個人間の遺伝的差異やがんゲノムの突然変異を高い正確さで同定する．
- アプローチ
  - ParaHaplo: Haplotype頻度を用いたType I errorの確率計算を，ハイブリッド並列で実装したマルコフ連鎖モンテカルロ法で行う．
  - ExRAT: RAT(Rapid Association Test)を，データ分割法で実装したインポータンスサンプリング法で行う．
  - NGS analyzer: ヒト標準ゲノム配列に対するマッピングと確率計算に基づいた多様性検出を，領域分割で実装した直接法による密度行列の対角化により行う．

現在までの研究開発成果

- 現時点の開発ソフトウェア ParaHaplo について
  - ParaHaplo ver. 3.0, phase II-2, 96%
  - 富士通コンパイラ（RICC，FX1，「京」，の全て）対応済，実行成功
  - PCクラスタで8,000ノードの並列性能を実証．
  - 「京」でも並列性能が出ることを確認．
  - 「京」ではコア並列（ハイブリッド並列）も実装している．
  - 現在は，一層の高速化を目指しSIMD化を行っている．
  - ソースコード公開済．
  - 論文出版：Misawa K, Kamatani N (2011) ParaHaplo 3.0: a program package for imputation and haplotype-based whole-genome association study using parallel computing. Source Code Biol Med.

現在までの研究開発成果

- 現時点の開発ソフトウェア ParaHaplo について（続き）
  - ParaHaploを実際のデータに応用した例.
  - 点線より上が従来の方法で遺伝子頻度の差が検出できた場所.
  - 実線より上がParaHaploで遺伝子頻度の差が検出できた場所.



現在までの研究開発成果

- 現時点の開発ソフトウェア ExRAT について
  - ExRAT ver. 1.0β- r382, phase II-2, 100%
  - 富士通コンパイラ（RICC, FX1,「京」, の全て）対応済, 実行成功, buildも可能.
  - PCクラスタで8,192ノードの並列性能を実証.
  - データ: 患者708人, 一般集団3397人, 8314SNP, 14100回permutationで評価.
  - ハイブリッド並列化対応: Preprocess処理とpermutation処理にOpenMPでのスレッド化の実装.
  - Preprocess処理の見直しによるさらなる高速化: 分割表探索処理での計算処理の見直しにより, 約40%程度高速化(計算処理を探索中に積み上げていけるように処理を変更し, 全体の計算量の削減に成功した).
  - MPIプロセスの増加によるメモリ使用量増加への対応: 非同期通信用のためのバッファ確保が原因だったため, 通信処理を見直し, 処理完了待ちのバッファの大きさを最小とすることで対処した.
  - テスト運用の結果から問題点の洗い出しおよび改善方法の検討を行ってきた.

現在までの研究開発成果

- 現時点の開発ソフトウェア ExRAT について（続き）
  - 検討の結果, 今後, 速度向上が期待される改善点:
    - スレッド数を増やしたときの処理効率の向上: スレッド間での同期待ちを出来るだけ少なくし, 処理効率向上を図る.
    - データサーバとの通信処理の改善: データサーバのレスポンスを改善するとともに, 通信方法も見直し, 全体の効率向上を進める.
    - MHサンプリングの処理で, スレッドを十分に活用できてないので, これを改善する方法を検討する.
    - 複数のSNPの組合せについての計算を, 1ジョブで処理できるように改良する.

Cores	Time (s)
16	12181
256	672
512	340
1024	173
2048	86
4096	58
8192	102

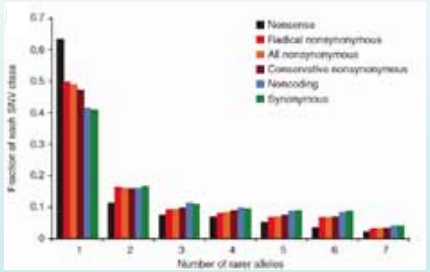
現在までの研究開発成果

- 現時点の開発ソフトウェア NGS analyzer について
  - NGS analyzer ver. 1.0, phase II-1, 80%
  - PCクラスタ上で2000並列を達成.
  - 全ゲノムシークエンスデータの解析
    - ヒトの全ゲノムシークエンス解析からの一塩基多型（SNP）, 挿入・欠失（indel）, リアレンジメントを高い精度で検出
    - がん細胞における突然変異を高い精度で検出（東京大学医科学研究所ヒトゲノム解析センターのPCクラスターにて実行）
  - 変異や多型のタイプごとに検出アルゴリズムを実装した。さらに、実験的検証を行うことで、パラメーターのチューニングを行った。
  - 初の日本人の全ゲノムシークエンス（*Nature Genetics* 2010）、肝臓がんの全ゲノムシークエンス（*Nature* 2010）を解析
  - Fujimoto *et al*. *Nature Genetics*, **42**, 931-936 (2010).
  - The International Cancer Genome Consortium. *Nature*, **464**, 993-998 (2010).

現在までの研究開発成果

- 現時点の開発ソフトウェア NGS analyzer について（続き）
  - 初の日本人の全ゲノムシークエンス（*Nature Genetics* 2010）の解析例：
    - ヒトの遺伝的多様性の例として1番染色体上のSNPの密度の計算結果.
    - パーソナルゲノム上に存在するSNPの頻度スペクトラムについての新たな知見.

ヒトの遺伝的多様性(1番染色体上のSNPの密度)



SNPの頻度スペクトラム

Fujimoto *et al. Nature Genetics*, **42**, 931-936 (2010).

プロジェクト終了時の達成目標

- 達成目標:
  - ParaHaplo: 全ゲノムの個人ごとの遺伝情報の違いの中から，疾患に関連する遺伝情報を網羅的に精度よく探し出すことを行うプログラムを開発し，「京」の上で運用できるように実装すること.
  - 第一走者 ParaHaplo で，ハプロタイプゲノムワイド関連解析を，数千人の疾患群とコントロール群とのハプロタイプ頻度の比較を行い，新たな疾患関連遺伝子の探索を行うことを目指す．8万ノードを使い，計算時間は1つの疾患あたり1時間～数時間程を予定している.
  - ExRAT: 遺伝子どうしが相互作用を起こして疾患への発症リスクを上昇する現象とそれらの遺伝子を新たに疾患関連遺伝子として発見するためのプログラムを開発すること．そして実際に新たな疾患関連遺伝子の探索を行うことを目指す.
  - NGS analyzer: 全ゲノム配列（約30億塩基）の個人ごとの遺伝情報の違いを網羅的に精度よく検出するプログラムを開発すること．そしてがんの全突然変異を高速に検出し，創薬のターゲット分子を探索することに資することを目指す.

大規模遺伝子ネットワーク推定ソフトウェア SiGN とデータ解析融合プラットフォーム Systems Biology integrative Pipeline (SBiP)

宮野　悟
次世代計算科学研究開発プログラム
データ解析融合研究開発チーム チームリーダー

発表者紹介

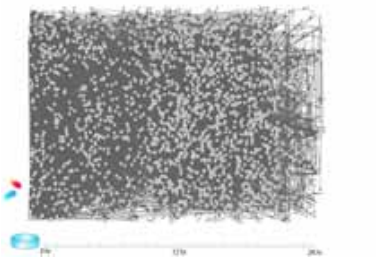
1977 年 3 月　九州大学理学部数学科卒業
1979 年 3 月　九州大学大学院理学研究科修士課程数学専攻修了
1979 年 6 月　九州大学理学部助手
1985 年 4 月　Alexander von Humboldt 財団研究員
1987 年 4 月　Universität GH Paderborn 助手
1987 年 12 月　九州大学理学部助教授
1993 年 3 月　九州大学理学部教授
1996 年 4 月　東京大学医科学研究所教授

研究分野

計算システム生物学、バイオインフォマティクス、メディカルインフォマティクス

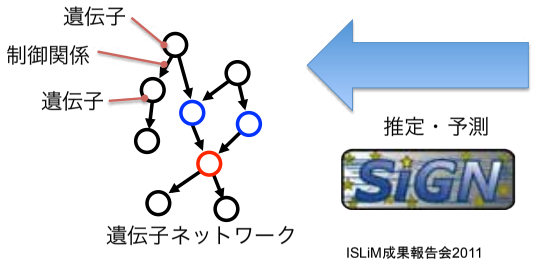
大規模遺伝子ネットワーク推定ソフトウェア SiGNとデータ解析融合プラットフォーム Systems Biology integrative Pipeline (SBiP)

データ解析融合研究開発チーム
東京大学医科学研究所
宮野　悟



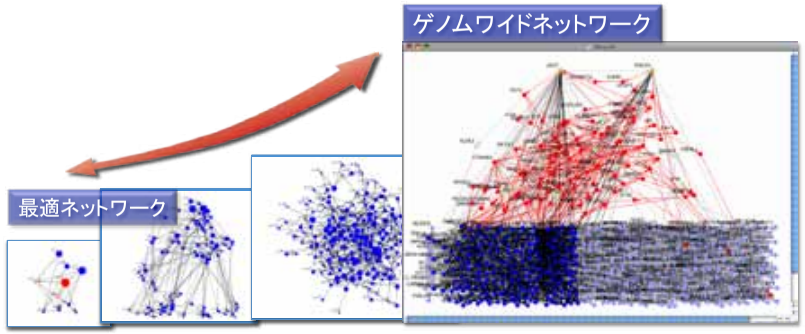
大規模遺伝子ネットワーク推定ソフトウェアSiGNの開発目的

肺がんなどの分子ネットワーク・薬剤応答ネットワークを解析し、その分子病態を描出するために、DNAチップ等で計測された遺伝子発現データから遺伝子間の発現の依存関係を表す「遺伝子ネットワーク」を推定・予測するソフトウェアを開発すること。



SiGNの目標

- 最適な数十遺伝子規模のネットワークからゲノムワイド（2万遺伝子）ネットワークまで、動的・静的ネットワークの推定が可能となること。



遺伝子発現プロファイルデータからの遺伝子ネットワークの抽出法

◇◆地図の作り方◆◇

- システムを捉えるための観測データ
  - 刺激後の時系列データ
  - 遺伝子KD後のデータ、など
- 因果・制御関係の数理モデル
  1. BN: Bayesianネットワーク＋非線形回帰
  2. SSM: 状態空間モデル＋次元圧縮
  3. “Modulator”と構造方程式を用いたグラフィカルモデル

SiGN+スパコンで、たとえば、次のようなことが可能になってきた。

- 抗がん剤に応答して動的に変化する大規模な遺伝子ネットワークを描出すること
- 数百のがんサンプルの遺伝子発現プロファイルデータから個々のがんの個性・多様性を分子ネットワークとして抽出すること
- これらを地図として用いて、新たな分子標的探索の可能性がでてきた

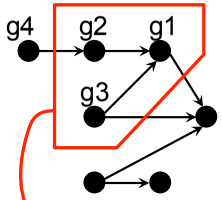
大規模遺伝子ネットワーク推定技術の現在までの主な成果

- SiGN-BN (ベイジアンネットワーク)
  - 大規模全ゲノム遺伝ネットワーク推定アルゴリズム(SiGN-BN-NNSR)
  - 中規模高効率並列ブートストラップアルゴリズム(SiGN-BN-HC+Bootstrap)
  - ➢ 京での並列化が大きく進行中
  - 小規模並列版静的・動的全体最適化アルゴリズム(SiGN-Para-OS)
  - ➢ 効率のよい高並列化は難しい（256ノードぐらいの長い計算が効率的）
    32ノードの最適ベイジアンネットワーク推定を達成（世界記録）
- SiGN-SSM (状態空間モデル)
  - 遺伝子ネットワーク推定アルゴリズム
  - モジュールネットワーク推定アルゴリズム
  - 遺伝子発現予測アルゴリズム
  - ➢ RICCで8192コア。京での並列化を準備中
- SiGN-L1 (L1正則化ネットワーク)
  - NetworkProfilerによる（京での並列化を準備中）
  - NetComparator による複数ネットワーク同時推定・比較アルゴリズム

SiGN-BN（ポスターD-4 ）

“仕事場における人間関係のようなもの”

Bayesian Network



- Markov assumption

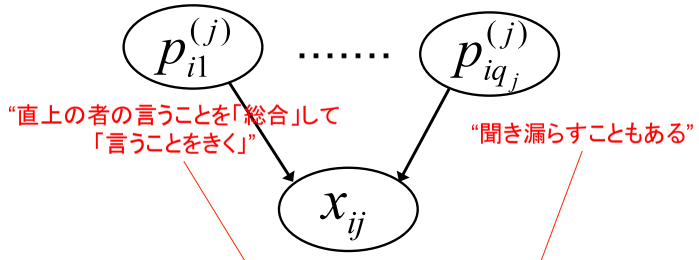
  「直上の言うことしか耳を傾けない」
- ネットワーク構造がわかっていれば、データがたくさん観測されると、

  「言うことをきく」確率がわかってくる。

$$x_{i1} \Leftarrow \mathbf{p}_{i1} = (x_{i2}, x_{i3})^T$$

$$f(x_{i1},...,x_{ip} \mid \boldsymbol{\theta}_G) = \Pi_{j=1}^{p} f_j(x_{ij} \mid \mathbf{p}_{ij}, \boldsymbol{\theta}_j)$$

Nonparametric Regression



We consider the additive regression model:

$$x_{ij} = m_1(p_{i1}^{(j)}) + \cdots + m_{q_j}(p_{iq_j}^{(j)}) + \varepsilon_j,$$
$$where \quad \varepsilon_j \sim N(0, \sigma_j^2) \; and \; \mathbf{p}_{ij} = (p_{i1}^{(j)},..., p_{iq_j}^{(j)}).$$

Here $m_k(\cdot)$ is a smooth function from R to R.

“数学的にはこんな感じになります”

Nonlinear Bayesian network model

$$f(x_{i1},\ldots,x_{ip};\boldsymbol{\theta}_G)=\prod_{j=1}^{p} f_j(x_{ij}\mid \mathbf{p}_{ij};\boldsymbol{\theta}_j),$$

$$f_j(x_{ij}\mid \mathbf{p}_{ij};\boldsymbol{\theta}_j)=\frac{1}{\sqrt{2\pi\sigma_j^2}}\exp\left\{-\frac{(x_{ij}-\mu_{ij})^2}{2\sigma_j^2}\right\}$$

$$\mu_{ij}=m_1(p_{i1}^{(j)})+\cdots+m_{q_j}(p_{iq_j}^{(j)})$$

$$=\sum_{k=1}^{q_j}\sum_{m=1}^{M_{jk}}\gamma_{mk}b_{mk}^{(j)}(p_{ik}^{(j)})$$

たくさんのパラメータ

“難しいのはみんなの仕事ぶりデータからネットワーク構造を推定すること”

Criterion for selecting good networks

BNRC Score

Bayesian Network and Nonparametric Regression Criterion

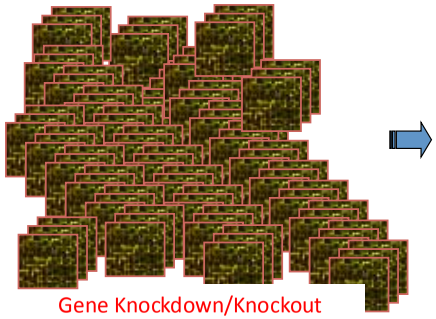
$$\mathrm{BNRC}(G)=-2\log\pi_G\int\prod_{i=1}^{n} f(\mathbf{x}_i;\boldsymbol{\theta}_G)\pi(\boldsymbol{\theta}_G\mid\boldsymbol{\lambda})d\boldsymbol{\theta}_G$$

$$=-2\log\pi_G-r\log(2\pi\, n^{-1})$$

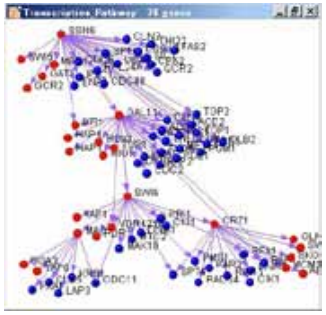
$$+\log\left|J_\lambda(\hat{\boldsymbol{\theta}}_G)\right|-2nl_\lambda(\hat{\boldsymbol{\theta}}_G\mid\mathbf{X}_n)$$

この”BNRC score”が小さいネットワーク構造とパラメータを探索する。“スパコンがここで必要になる”

Bayesian Network + Nonparametric Regression
Dynamic BN（時系列データに限定した方法）



Gene Knockdown/Knockout
Time-Course Mesurement



遺伝子ネットワーク

1. Imoto, S., Goto, T., Miyano, S. Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression. Pacific Symposium on Biocomputing. 7:175-186, 2002.
2. Imoto, Kim, Goto, Aburatani, Tashiro, Kuhara, Miyano (2003). Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic networkJ. *Bioinformatics and Comp. Biol.*, 1(2), 231-252

メラノーマ細胞の遺伝子ネットワークとタキソール

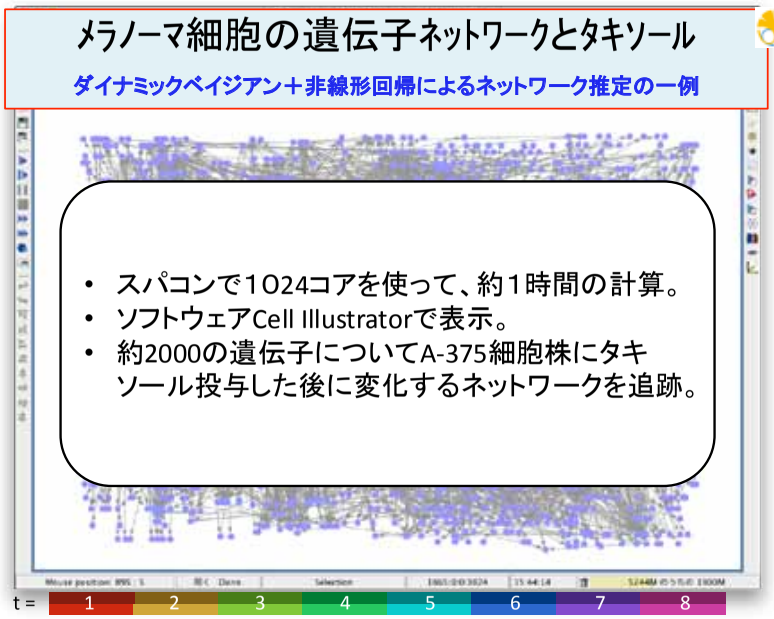
具体的にどんなことが今の段階でできているのか－ダイナミックベイジアン＋非線形回帰

- 細胞：メラノーマ細胞株 A-375
- 薬：タキソール(Paclitaxel、抗がん剤)
- Microarray：Illumina® HumanHT-12 v3
- データ：14 時点、3回計測、計84アレイデータ
  - 0h, 15min, 30min, 45min, 1h, 1.5h, 2h, 3h, 4h, 6h, 8h, 12h, 18h, 24h
  - タキソール投与と非投与の2種類のデータ
- 目的：タキソールを投与したメラノーマ細胞の遺伝子ネットワークの動的変化を見ることで、新たなターゲット遺伝子を探索

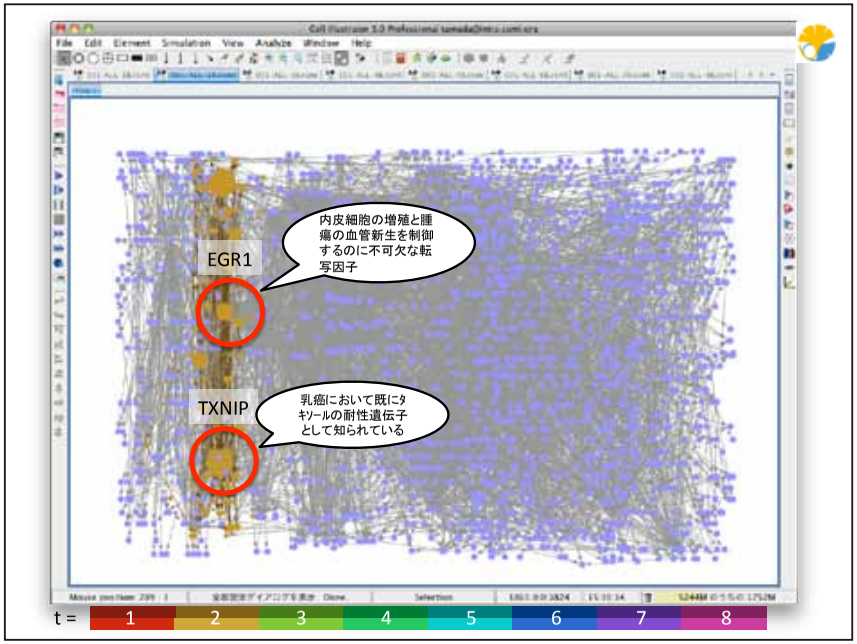
Cell Innovator, C. Print (U Auckland)との共同研究

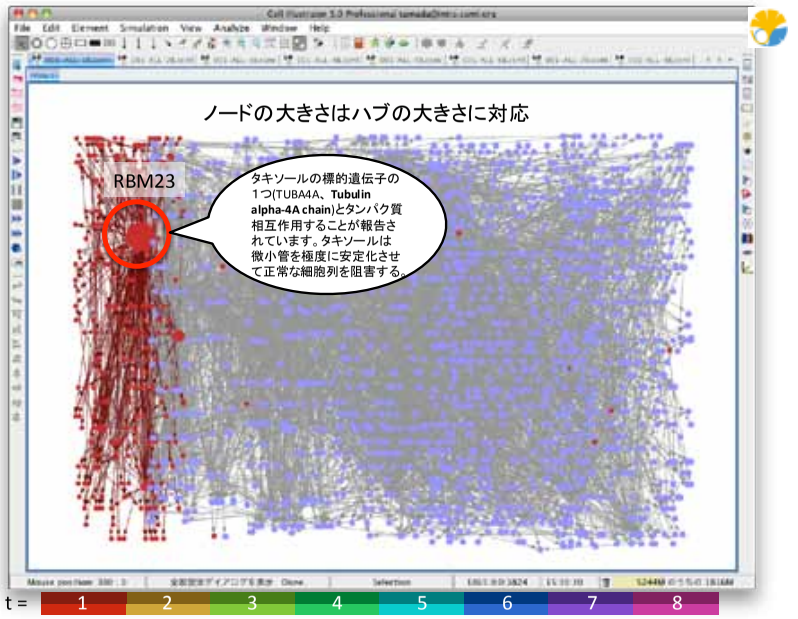
メラノーマ細胞の遺伝子ネットワークとタキソール

ダイナミックベイジアン＋非線形回帰によるネットワーク推定の一例



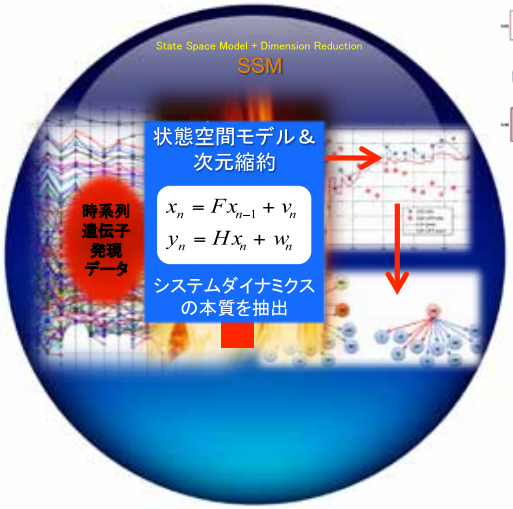
- スパコンで1024コアを使って、約1時間の計算。
- ソフトウェアCell Illustratorで表示。
- 約2000の遺伝子についてA-375細胞株にタキソール投与した後に変化するネットワークを追跡。

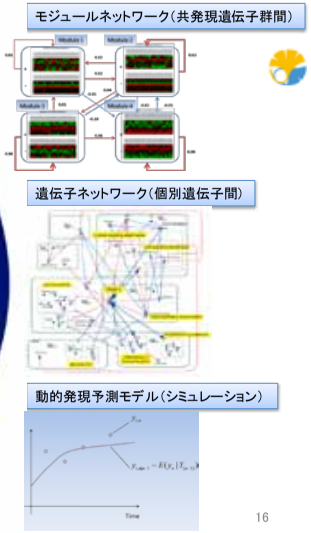




状態空間モデルによる予測

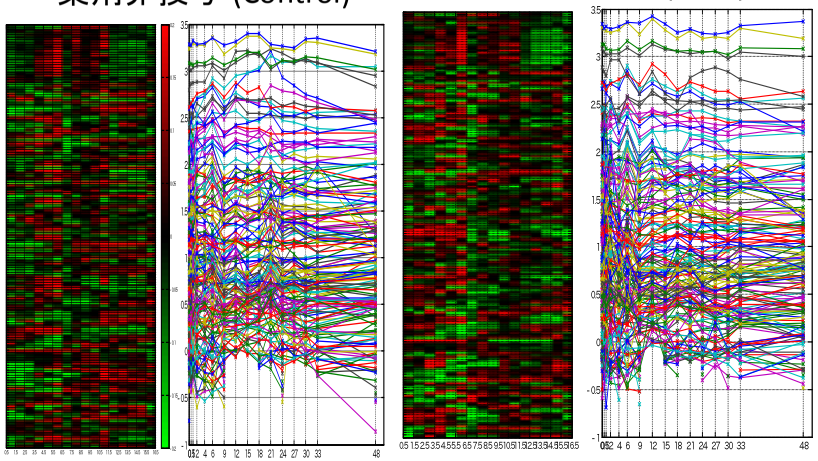
SiGN-SSM（ポスターD-6）





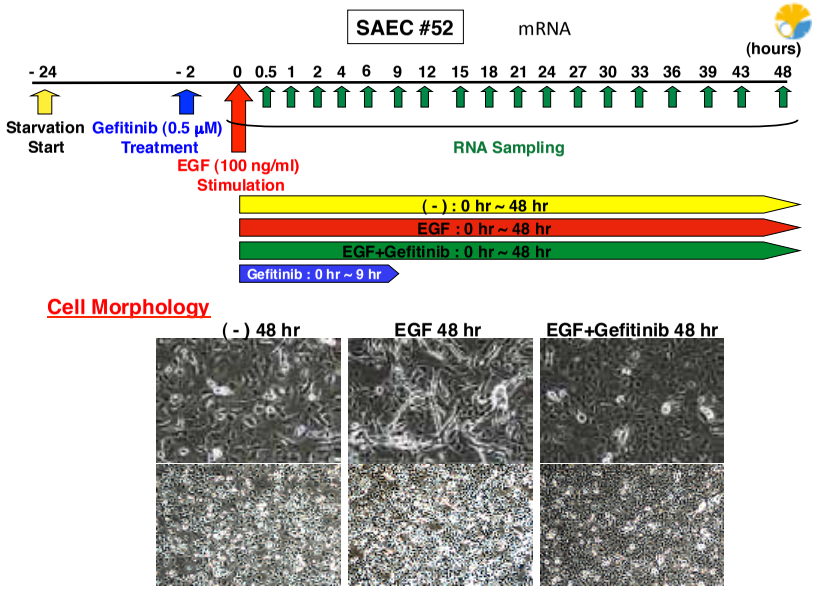
背景：時系列遺伝子発現データ

薬剤非投与 (Control)　　薬剤投与 (Case)



背後にある生体システムの差異を知りたい

17



正常肺上皮細胞SAEC＋EGF(+Gefitinib) 投与下 EGF 応答パスウェイ解析

■ 使用データ
細胞：肺正常細胞 SAEC
刺激 / 薬：EGF，Gefitinib(イレッサ：抗がん剤)
マイクロアレイ：Agilent 44K Whole Human Genome Oligo Microarray
計測データ：19 時点 × 1 回計測 in 48 時間
EGF,Gefitinib 投与・非投与の 4 条件

■ 目的
Gefitinib の影響により遺伝子間制御構造に変化のあった遺伝子群の予測および，肺がん予後予測バイオマーカ探索

■ 利用手法
コントロールデータにより構築されたモデルからのケースデータの予測に基づくシステム異常検知

■ 結果　遺伝子間制御構造に差異のある遺伝子群の予測および肺がん予後予測シグネチャーへの応用に成功

動的発現予測　　発現制御構造推定　　制御構造に差異のある遺伝子群の発現パターンに基づく肺がん患者群予後予測

ISLiM成果報告会2011

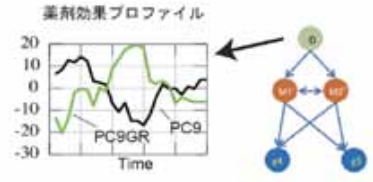
19

肺がん細胞株PC9/PC9GR に対するGegitinib 動的効果推定

■ 使用データ
細胞：肺がん細胞株 PC9 および PC9GR
刺激 / 薬：EGF，Gefitinib(イレッサ：抗がん剤)
マイクロアレイ：Agilent 44K Whole Human Genome Oligo Microarray
計測データ：26 時点 × 1 回計測 in 26 時間

■ 目的
薬剤耐性能の異なるがん細胞への薬剤作用点および作用機序の推定：
PC9（薬剤感受性）PC9GR（薬剤耐性）

■ 利用手法
薬剤動的効果モデル項を組み込んだモデルによる薬剤作用下データの予測に基づく状態推定



PC9 と PC9GR の間で異なる薬剤効果プロファイルを持つモジュールを同定．
当該モジュール支配下にある遺伝子群が薬剤耐性能の差異に関わる可能性を示唆．

ISLiM成果報告会2011

20

SiGN-SSM（ポスターD-6 ）

L1正則化法による大規模遺伝子ネットワーク推定プログラム

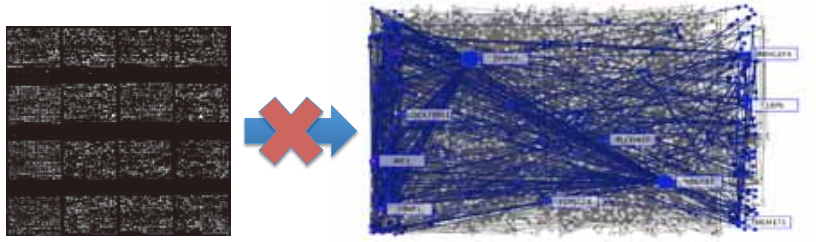
- SiGN-L1 は L1 正則化法によるスパースな統計的グラフィカルモデル（グラフィカルガウシアンモデル、ベクトル自己回帰モデル、構造方程式モデル）を用いた遺伝子ネットワーク推定ソフトウェア

SiGN-L1でできたこと

1サンプルから因果・制御のネットワークを抽出するのは無理

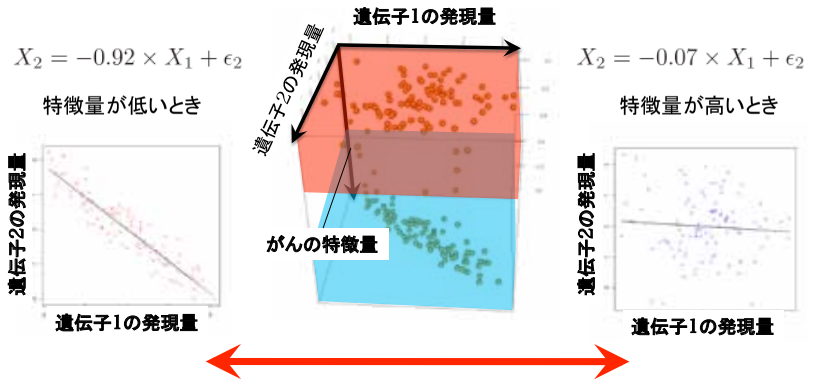
クラスタリング
2サンプルだとFold-Change Analysis

1サンプル　　因果・制御のネットワーク



しかし、たくさんのサンプルがあると・・・

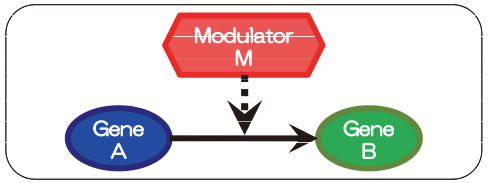
データの層別化による相関構造の違い



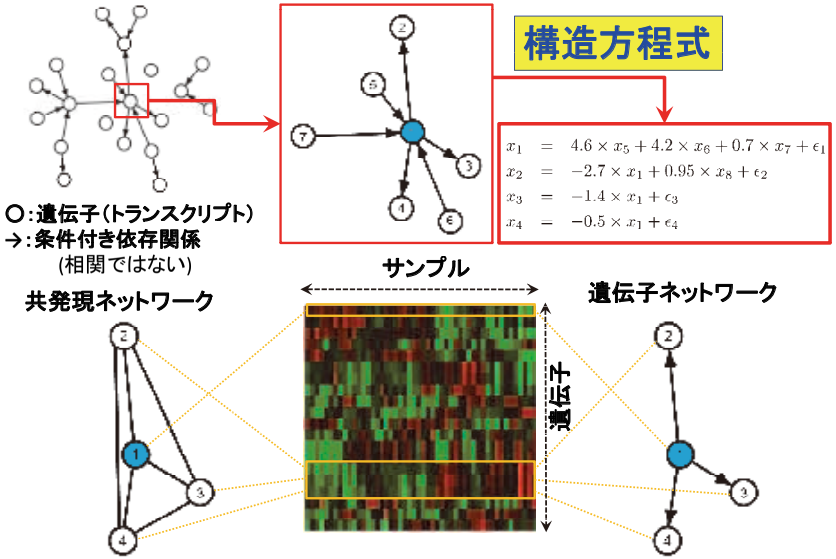
データの層別化

がんの「特徴量」を導入すれば、関係がでてくる
⇒サンプルそれぞれのネットワークを推定できる。

モジュレータ: 遺伝子AとBの条件付き従属性に影響を与える因子を導入



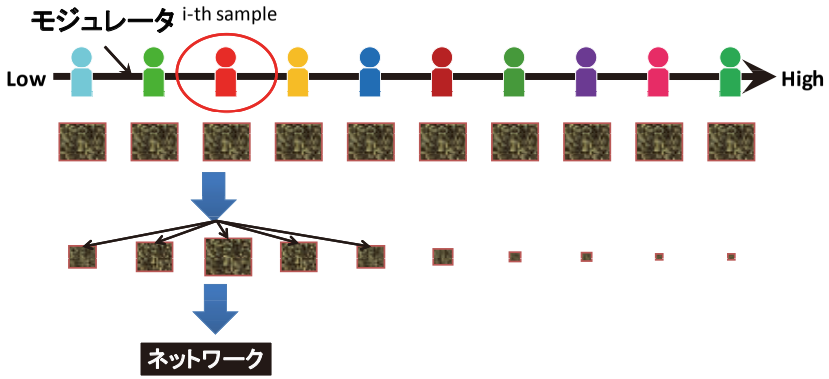
構造方程式モデル



NetworkProfiler

i番目のサンプルの構造方程式モデルの係数を推定

Sample Weightingを使う



がんの個性を捉えるためのネットワーク推定法

モジュレータによる特徴空間にサンプルを射影し、
サンプルそれぞれの遺伝子ネットワークを推定する

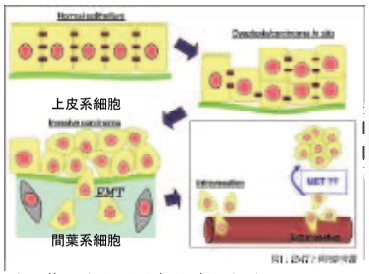
がんの特徴ベクトル（モジュレータ）

スパコンで捉えたがんの黒幕たちのネットワーク

“上皮間葉を引き起こす遺伝子の発見”
グランドチャレンジでの大規模データ解析の事例

上皮間葉転換(Epithelial-Mesenchymal Transition **EMT**)

- 上皮系細胞が間葉系細胞に形態変化する現象
- がんの悪性度、がんの浸潤性、線維症に深く関連
- ゲフィチニブなどの抗がん剤を効かなくする原因



?
EMTを引き起こすいくつかの遺伝子は見つかっているが、そのメカニズムはよくわかっていない

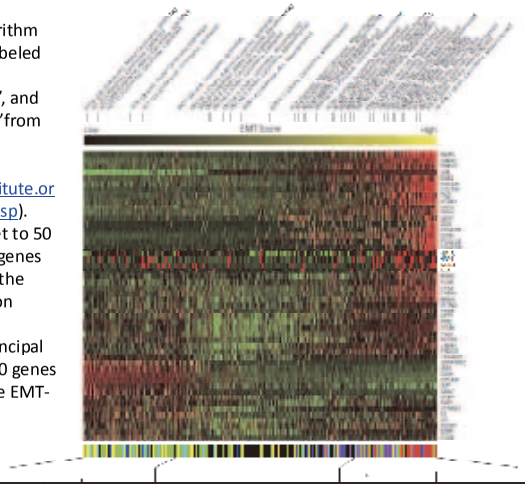
http://gantoku3.umin.jp/topics/saito.html

モジュレータ：EMTness（低：上皮細胞，高：間葉細胞）

50 個の上皮･間葉マーカー遺伝子の第一主成分

Signature-based hidden modulator extraction algorithm

1. Selected 122 genes labeled “EMT UP”, “EMT DN”, “JECHLINGER EMT UP”, and “JECHLINGER EMT DN”from Molecular Signatures Database v2.5 ([6]; http://www.broadinstitute.org/gsea/msigdb/index.jsp).
2. Then, narrowed the set to 50 functionally coherent genes with $p < 10^{-5}$ by using the extraction of expression module (EEM) .
3. Computed the first principal component of these 50 genes as hidden values of the EMT-related modulator



EMT に関わる がん細胞株の遺伝子ネットワークの推定

Sanger Center 公開データ
762 個のがん細胞株マイクロアレイデータ
様々ながん種
13,508 transcripts=13,006 mRNA+502 miRNA

モジュレータ：EMTness（低：上皮細胞，高：間葉細胞）

1,024 cores (12.3 TFLOPS)を使って3ヶ月の計算

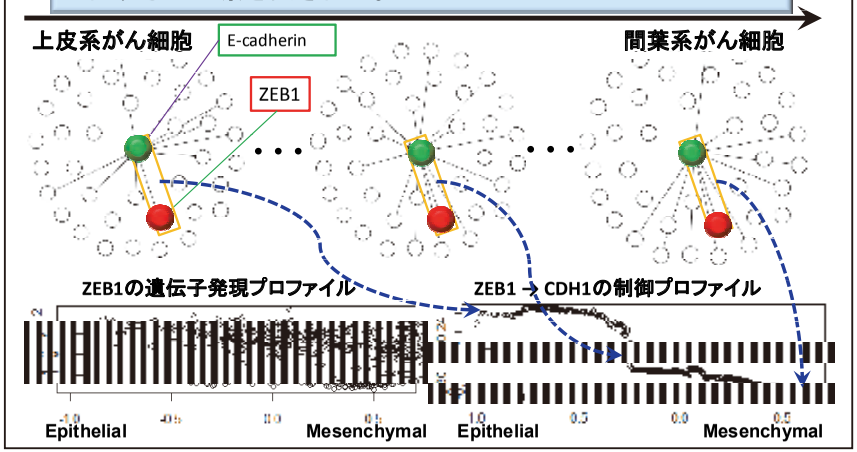
EMT に関わる 762 がん細胞株の遺伝子ネットワーク

ISLiM成果報告会2011

グランドチャレンジでの大規模データ解析の事例

EMTで変化する大規模遺伝子ネットワークを1024コアで3ヶ月間かけて計算

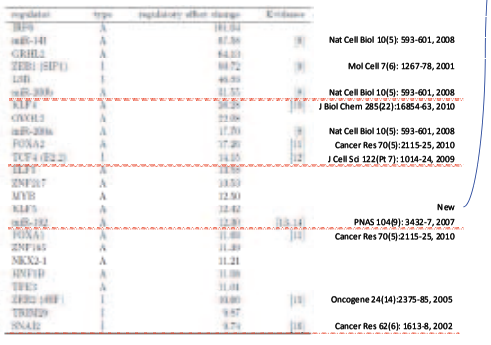
- 13,508個の遺伝子から構成される遺伝子ネットワークを762個のネットワークが EMTの度合でその構造がどのように変化するかが暴き出された。



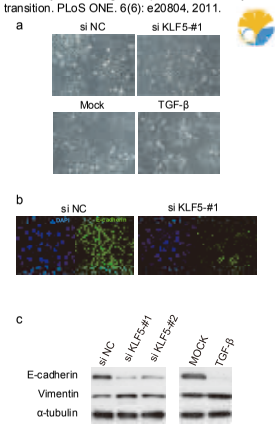
スパコンが予測したEMT制御因子たちと新規遺伝子KLF5の発見

スパコンによる大規模データ解析の有効性

- 名古屋大学医学研究科の高橋隆教授らのグループが、遺伝子KLF5をノックダウンするとEMTが引き起こされることを肺がん細胞株で実証。
- スパコンがEMT制御因子として予測したトップ24遺伝子のうち11遺伝子が「当たっていた」。

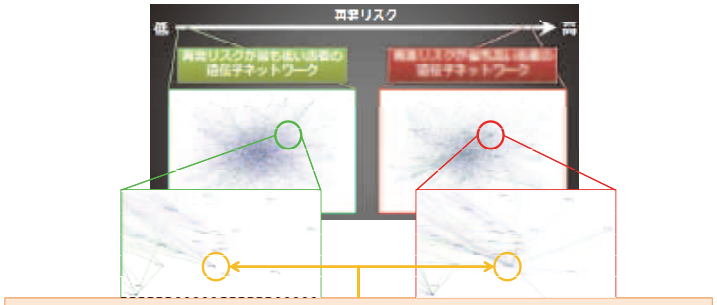


Shimamura T, Imoto S, Shimada Y, Hosono Y, Niida A, Nagasaki M, Yamaguchi R, Takahashi T, Miyano S. A novel network profiling analysis reveals system changes in epithelial-mesenchymal transition. PLoS ONE. 6(6): e20804. 2011.



再発リスクに関わる日本人肺癌患者の遺伝子ネットワークの比較

データ: 226症例の肺線がん患者の遺伝子発現データ
（国立がん研究センター横田淳先生、河野隆志先生との共同研究）
目的: 再発リスクをモジュレーターとしたときの 226 症例各々の遺伝子ネットワークを推定し、再発リスクが高い患者と低い患者のシステムの違いを比較

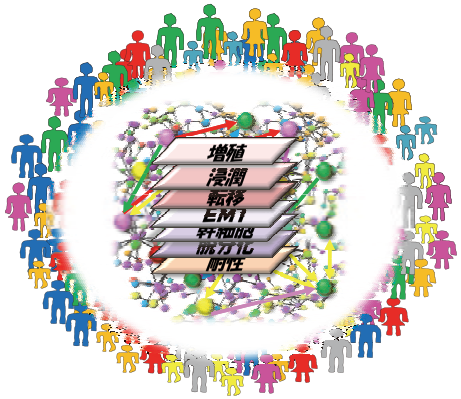


CTGF (Connective Tissue Growth Factor)
増殖・分化を制御するTGF-βの下流因子で、TGF-βの間質繊維化促進作用を仲介
増殖促進、遊走、細胞外基質算出、血管新生作用を呈する

33

500種類の解析するには どれだけ計算リソースが必要か？

1000コア（12TFLOPS)だと25年
K computerだと2週間



世界水準との比較ーSiGN

- 京に代表される超並列計算機を用いる遺伝子ネットワーク推定プログラムパッケージは現時点ではSiGNを除いて世界に無い。
- よく知られているARACNE　(Andrea Califano, Columbia Genome Center) とは、精度・解析規模・解析法の多様性の点からSiGNが優れている。

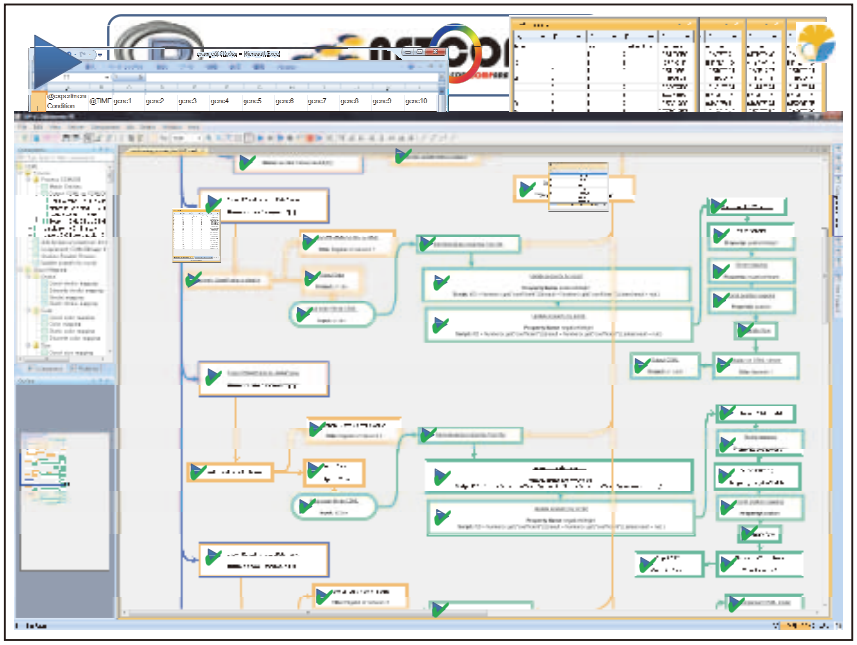
ISLiM成果報告会2011　35

データ解析融合プラットフォーム

SBiP (Systems Biology integrative Pipeline)

- SiGN、LiSDASを解析パイプラインのコンポーネントとしてユーザが簡単に利用できる高機能GUI を備えたソフトウェアプラットフォーム
- 京に用意されているジョブスケジューリングシステムと連携し、京にて各解析パイプラインの処理の一部を実行し、その結果を、SBiPの視覚化コンポーネント群を用いて保存できるよう目指している。
- ユーザはSBiPに用意されているさまざまな解析コンポーネント を組み合わせてカスタマイズした解析フローを実行可能。

ISLiM成果報告会2011　36



プロジェクト終了時の達成目標

- SiGNの高並列化（20000ノード　40000ノード）
- ネットワーク解析による薬のターゲット探索やがんなどの病態を理解する基盤技術の創出

ISLiM成果報告会2011

網羅的タンパク質間相互作用予測ソフトウェア MEGADOCK の開発と応用

秋山　泰
東京工業大学 大学院
情報理工学研究科計算工学専攻

発表者紹介

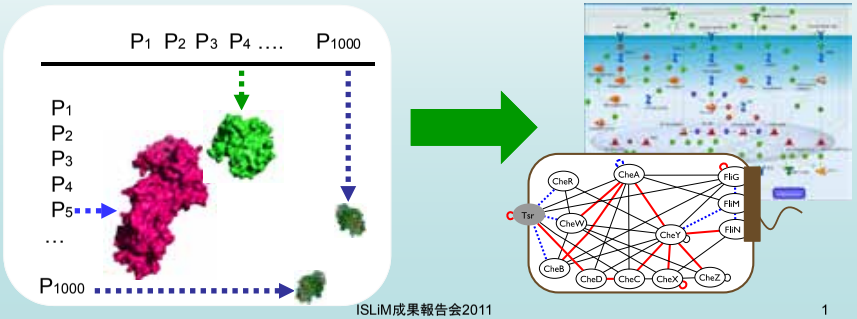
1990 年 3 月　慶應義塾大学大学院理工学研究科電気工学専攻博士課程修了(工学博士)
1990 年 4 月　通商産業省工業技術院電子技術総合研究所　研究官
1992 年 4 月　京都大学　助教授（化学研究所）
1996 年 4 月　技術研究組合新情報処理開発機構　研究室長（並列応用つくば研究室）
2000 年 4 月　通商産業省工業技術院電子技術総合研究所　主任研究官　（生命情報科学研究センター検討チーム長）
2001 年 4 月　独立行政法人産業技術総合研究所　生命情報科学研究センター長
2007 年 4 月　東京工業大学大学院情報理工学研究科計算工学専攻 教授

研究分野

バイオインフォマティクス、並列処理応用

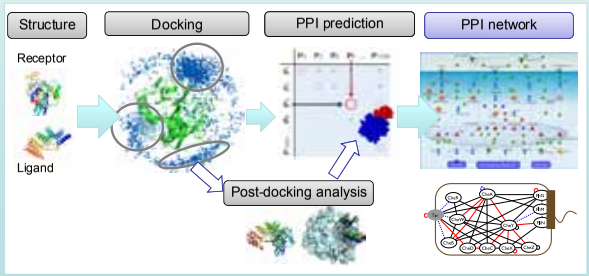
網羅的タンパク質間相互作用予測ソフトウェア MEGADOCKの開発と応用

東京工業大学
大学院情報理工学研究科 計算工学専攻
秋山 泰



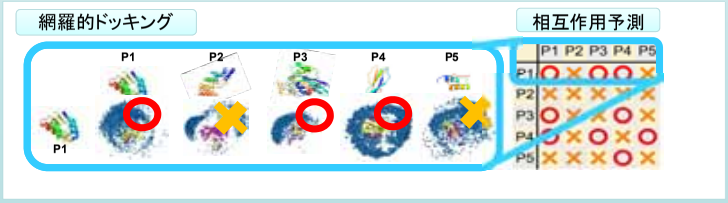
タンパク質間相互作用ネットワークの推定とその応用に関する研究

- 目的
  - タンパク質ネットワーク予測のための超大規模計算システムを開発する。
  - 膨大なタンパク質候補の中からタンパク質間相互作用を予測し、創薬ターゲットとして提示。個人の遺伝子型から表現型を予測するパイプラインの一部となる。
- 2012年までの目標
  - システム生物学が対象とする網羅的超大規模計算(2000x2000級)を現実的な時間で計算可能にする



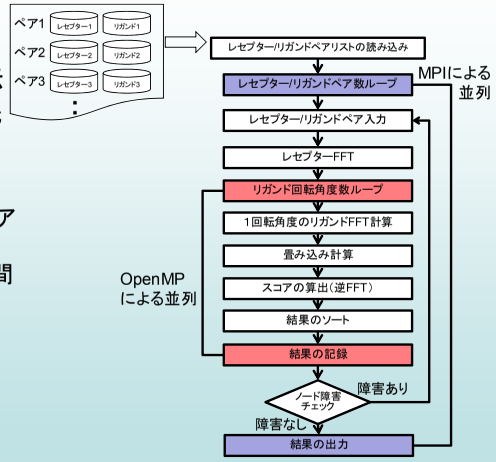
概要・アプローチ

- 研究開発コードの概要
  - タンパク質立体構造情報に基づき、表面形状相補性と静電相互作用を用いた単純化した評価モデルを新規に提案した。FFTを用いて計算量を大きく減じ、さらにOpenMPとMPIによるハイブリッド並列機能を実装することにより、大規模並列計算機上で効率的な並列計算を行う。
- MEGADOCKによるPPIネットワーク予測
  - 入力したタンパク質群の構造について、網羅的に剛体ドッキングを行い、複数の複合体候補構造を出力する。ポストドッキング解析により、タンパク質間相互作用ネットワークを予測する。



アプリケーションの概要

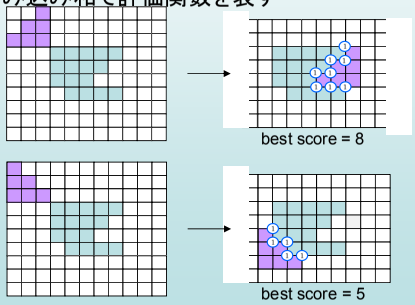
- 離散化(計算モデル化)の方法
  - 二つの構造間の結合性を三次元複素畳込みで評価
- 計算方法
  - 形状相補性と静電相互作用スコアを一つの複素数で表し(rPSC)、二体間の畳込み和をフーリエ空間上で行い逆FFTで評価値を得る
- 並列化の方法
  - ノード内の回転角度並列: OpenMP
  - ノード間のデータ並列: MPI



剛体ドッキング

- Rigid Docking (Grid base & Scoring function)
  - タンパク質立体構造の形状相補性等に基づいて評価
    - タンパク質をボクセル空間で表現
    - 特定のスコアリングモデルに基づいてスコアを付与
    - それらの畳み込み和で評価関数を表す

15度刻み
3600通りの探索



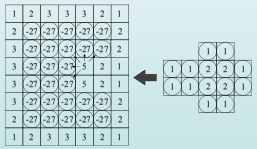
Katchalski-Katzir E, et al. *PNAS*, 1992.

新規に提案したスコアリングモデル: rPSC

- real Pairwise Shape Complementarity (**rPSC**)
  - 形状相補性のスコアを実数のみで表現

$$G_R(l,m,n)=\begin{cases}\#\text{ of R atoms within }(3.6\mathring{A}+\text{R atom }r_{vdw})\\ -27\ \text{ inside of the R}\end{cases}$$

$$G_L(l,m,n)=\begin{cases}0 & \text{solvent accessible surface layer of the L}\\ 1 & \text{solvent excluding surface layer of the L}\\ 2 & \text{core of the L}\\ 0 & \text{open space}\end{cases}$$



Receptor rPSC　Ligand rPSC

MEGADOCK Score
= rPSC + *i* Elec

$$R(l,m,n)=G_R(l,m,n)+iE_R(l,m,n)$$
$$L(l,m,n)=G_L(l,m,n)+iwE_L(l,m,n)$$
$$S(\alpha,\beta,\gamma)=\Re\left[\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N}R(l,m,n)L(l+\alpha m+\beta n+\gamma)\right]$$
$$S(\alpha,\beta,\gamma)=\mathrm{IFT}[\mathrm{DFT}[R(l,m,n)]^{*}\cdot\mathrm{DFT}[L(l,m,n)]]$$

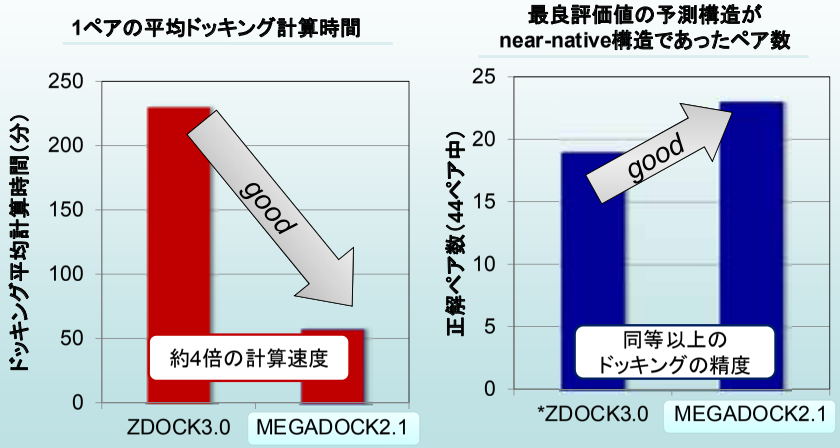
ZDOCK

PSC	=	Re	+*i*	Im
Elec	=	Re		
Desol	=	Re	+*i*	Im

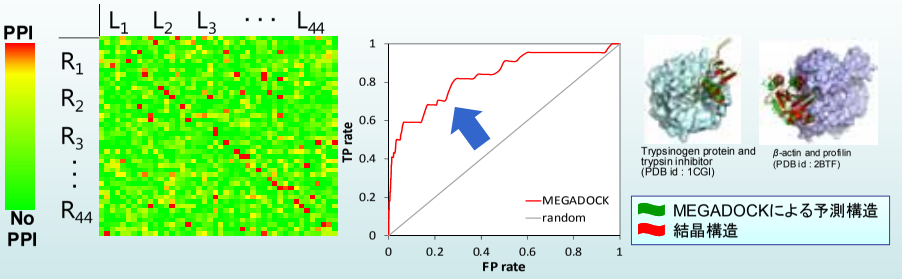
※既存モデル: **PSC** (ZDOCK*のモデル)
形状相補性スコアを複素数で表現
*Chen R, Weng Z. *Proteins*, 2003.

既存アプリケーションとの性能比較

44タンパク質ペア(PPD Benchmark 2.0)に対する実験結果



ベンチマークデータ (44x44=1936) による網羅的 PPI 予測結果

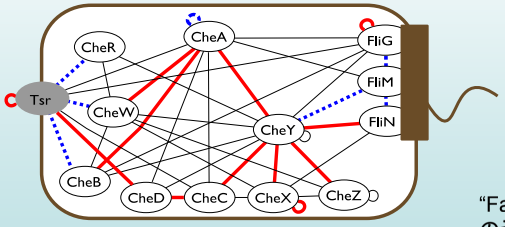


***F*-measure : 0.41**
より詳細なスコア関数のZDOCKを用いた予測精度(0.43)と同等レベル

$$F\text{-measure}=\frac{2\cdot\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})+(\mathrm{TP}+\mathrm{FN})}$$

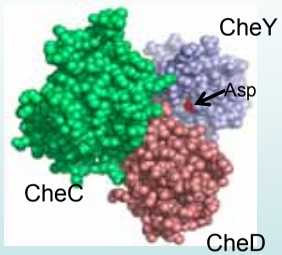
システム生物学の実問題への応用

- 細菌走化性シグナル伝達系
  - システム生物学における典型的問題
  - 89構造を用いた89レセプター対89リガンド =7921通りの計算



赤太線:True Positive
青点線:False Negative
黒実線:False Positive

***F*-measure : 0.44**

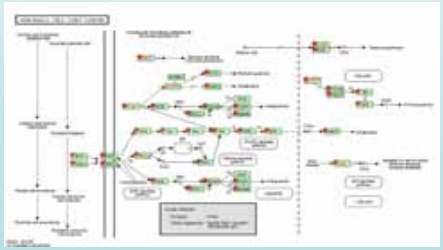


"False positive" と評価した相互作用のうちCheY-CheDなど未知相互作用の提案を検討

EGFRシグナル伝達系既知パスウェイの再構築による評価

- 超大規模解析(約25万件)の解析を完了
  - 497レセプター対497リガンド =247,009構造ペア
- 高い精度を示した
  - Precision (適合率) 0.29
  - Recall (再現率) 0.47
  - F-measure 0.36

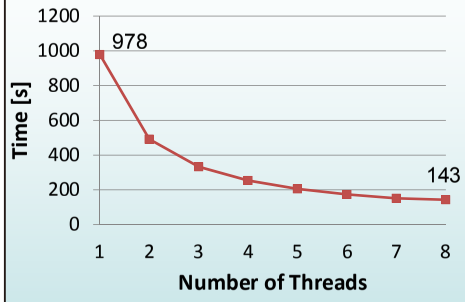
25万規模の大規模な実問題への適用においても、細菌走化性系(1万規模)での結果と同等の精度を得た

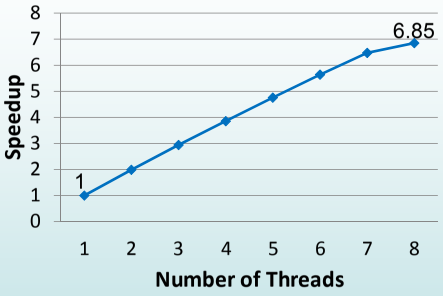


	正解ペア	不正解ペア
Positive判定	53	131
Negative判定	59	747

同じタンパク質のデータを縮約

「京」1ノードでのスケーラビリティ (スレッド並列)





8スレッド実行時に1スレッド実行時の**約6.85倍**の速度

FFT N = 108, PDB-ID: 1ACB のドッキング (30回実行した平均値)

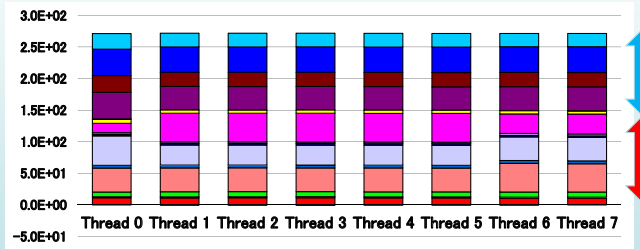
＊計測値は整備中のシステムによる暫定的な数値です。(2011年9, 11月計測)

今後のチューニング課題

ドッキング計算部分について詳細プロファイルを取得

計算時間のかかっている部分

バリア同期待ち　浮動小数点演算待ち　浮動小数点ロードキャッシュアクセス待ち

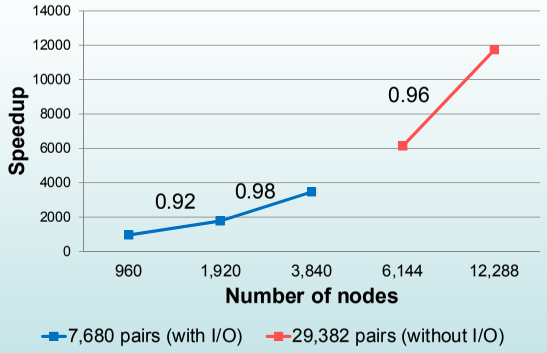


SIMD化率の向上にむけて

三つのFFTライブラリを比較する
(1)FFTW (現在)、(2)FFTE、(3)Conv3D
ただし Conv3Dはライセンス上別環境で測定

＊計測値は整備中のシステムによる暫定的な数値です。(2011年11月計測)

「京」におけるストロングスケーリング（OpenMP, MPI ハイブリッド並列）



スケーラビリティ
（6114→12288ノード）

$$\alpha = \frac{\frac{T_{12288}}{12288}}{\frac{T_{6144}}{6144}} = 0.96$$

T: 計算時間

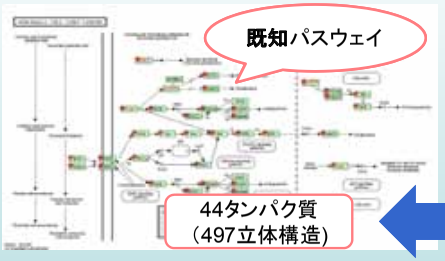
12,288 ノード x 8 コア ： 98,304コア並列を達成

FFT N = 140付近のタンパク質群約36,000件のドッキングを12,288ノード使用して行った場合、約17分で完了

＊計測値は整備中のシステムによる暫定的な数値です。（2011年9, 11月計測）

10Petaで挑む応用課題 肺がん関連パスウェイと薬剤関連タンパク間の新規相互作用探索

肺がん関連（EGFR系）の遺伝子ネットワークに属するタンパク質群を対象とした網羅的PPI予測



肺がん薬関連遺伝子（東大宮野研究室提供）に関するタンパク質データ

294タンパク質
（約1500 立体構造）

約2000x2000 級のドッキング計算結果を実施
新規相互作用の発見を目指す

80,000ノード利用の場合
全計算を約半日で終了する見込み
1/4時間 x 2,000 x 2,000 / 80,000ノード = 12.5 時間

ISLiM成果報告会2011

謝　辞

- 結果の一部は、理化学研究所が実施している京速コンピュータ「京」の試験利用によるものです。
- 「京」における計測値は整備中のシステムによる暫定的な数値です。（2011年6, 9, 11月計測）

ISLiM成果報告会2011　15

LiSDAS:データ同化計算技術に基づく生体情報シミュレーション

樋口知之
次世代計算科学研究開発プログラム
データ解析融合研究開発チーム

発表者紹介

1989 年　3 月　東京大学理学系研究科地球物理学博士課程修了
1989 年　4 月　統計数理研究所予測制御研究系予測理論研究部門助手
1994 年 12 月　統計数理研究所予測制御研究系予測理論研究部門助教授
2002 年　7 月　統計数理研究所予測制御研究系システム解析研究部門教授
2004 年　4 月　統計数理研究所モデリング研究系教授，予測発見戦略研究センター副所長（兼務）（～2010.3）
2011 年　4 月　統計数理研究所　所長

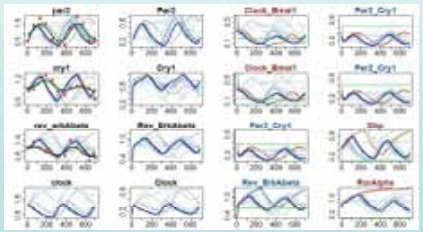
研究分野

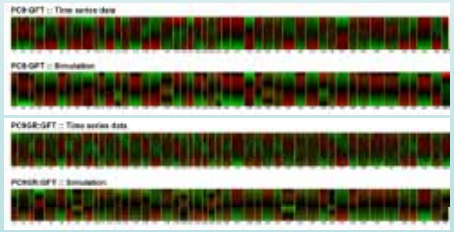
ベイジアンモデリング，データ同化

LiSDAS:データ同化計算技術に基づく生体情報シミュレーション

統計数理研究所・所長
樋口知之

統計数理研究所・データ同化研究開発センター
吉田亮　中野慎也　長尾大道　斎藤正也





I. ソフト研究開発の背景・目的

モデリングの制約

- タンパク質相互作用や転写因子データベースにもとづく反応経路列挙型モデリングの限界（組み合わせ爆発）
- モデリングの対象である分子リストが不完全なため、閉じた系として記述できない

実験の制約

- 細胞の環境依存性や生体内分子のバリエーションならびに個体間の遺伝的差異等がもたらす生化学反応の多様性
- 個体ごとに異なる生化学パラメータやネットワークトポロジー

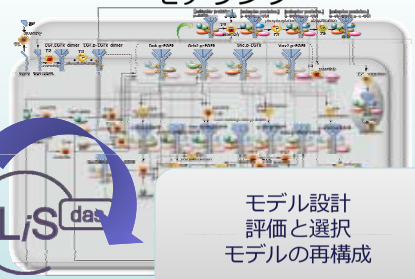
実験とモデリングの知識循環から永続的なモデル改良が必要である

LiSDASが提供する機能

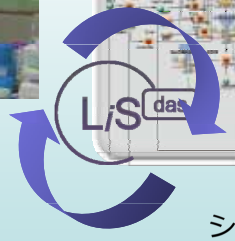
生体内分子の計測データを参照値としてあたえることで、シミュレーションの再現性・予測力を改善するためのパラメータチューニングやモデルの改良を自動で行う。

実験・計測

モデリング

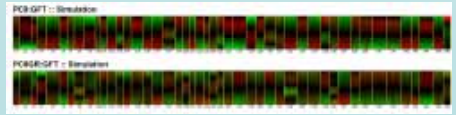


仮説構築
薬剤応答性の予測
実験のデザイン



モデル設計
評価と選択
モデルの再構成

シミュレーション



II. ソフト研究開発の概要・ アプローチ

データ同化によってパスウェイの構造、反応速度や状態変数の初期条件を推定する

状態空間モデル:

$$x_t = f(x_{t-1}, v_t \mid \theta)$$ 確率的シミュレーションモデル

$$y_t = h(x_t, w_t \mid \theta)$$ 観測モデル

θ: モデル・パラメータ

x_t: シミュレーション変数　y_t: 観測データ　v_t, w_t: エラー

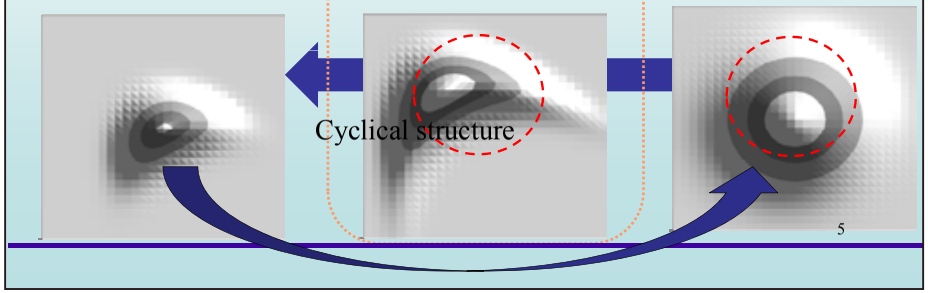
– 特に、生命科学で扱うデータでは$\dim(y_t) \ll \dim(x_t)$ であり、モデルパラメータの推定は一般には困難である。

4

ベイズ推定

$$p(\theta \mid y_{1:T}) \propto p(y_{1:T} \mid \theta) \cdot p(\theta)$$

Posterior	Likelihood	Prior
Improved knowledge about values of *x*	Feasibility of realization of *y* for given *x*	Belief about values of *x*



計算手法

ディリクレ混合過程で事前分布を設計して、MCMCで状態空間モデルのパラメータを推定する

基底測度 $G_0(\pi)$
観測時系列を分析して設計する

サンプリング π_i

サンプリング $\theta_j|\pi_i$

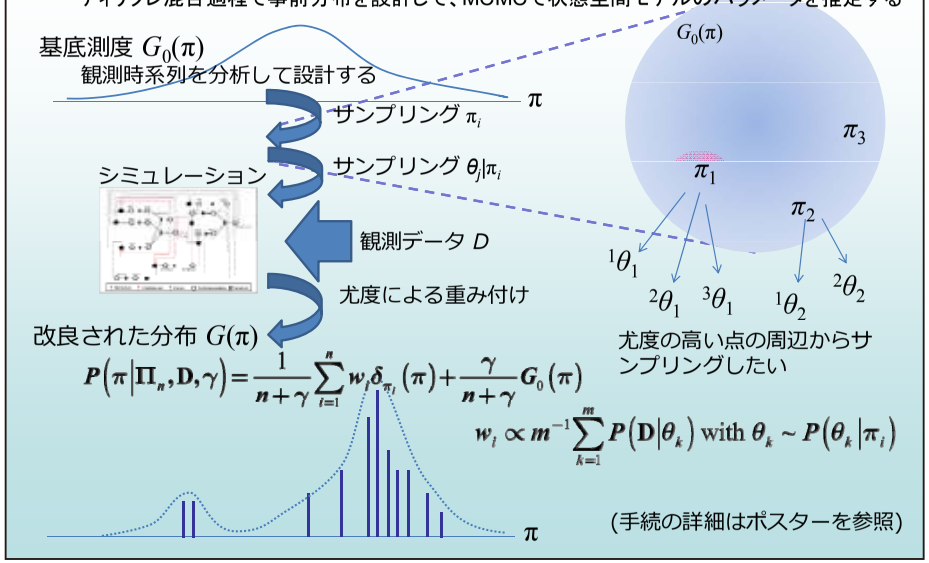
観測データ D

尤度による重み付け

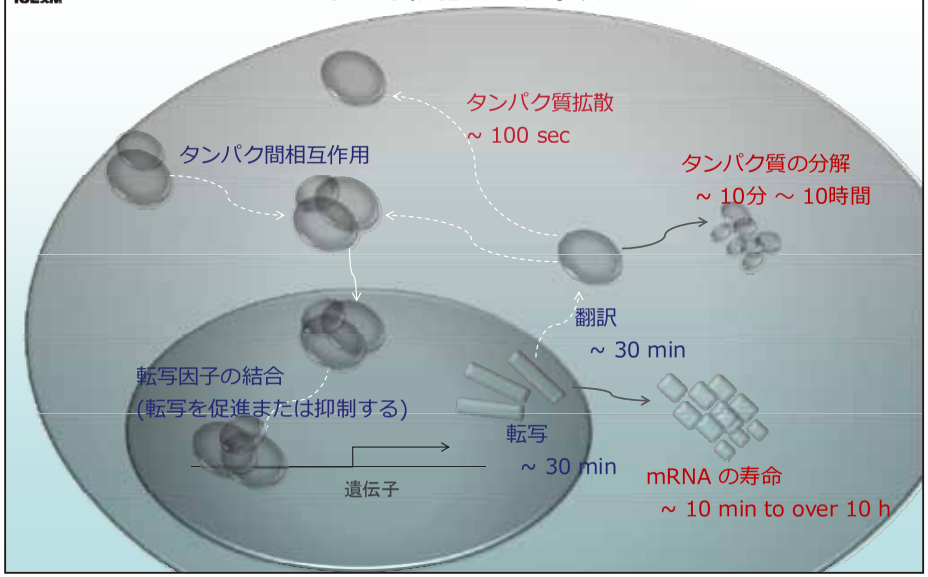
改良された分布 $G(\pi)$

$$P\left(\pi \middle| \Pi_n, \mathbf{D}, \gamma\right) = \frac{1}{n+\gamma}\sum_{i=1}^{n} w_i \delta_{\pi_i}(\pi) + \frac{\gamma}{n+\gamma} G_0(\pi)$$

$$w_i \propto m^{-1}\sum_{k=1}^{m} P\left(\mathbf{D} \middle| \theta_k\right) \text{ with } \theta_k \sim P\left(\theta_k \middle| \pi_i\right)$$



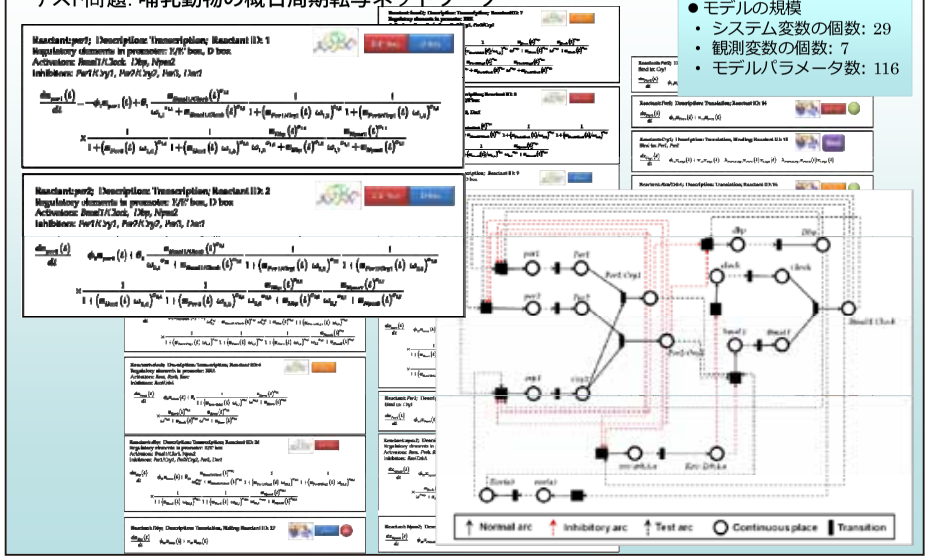
LiSDASにおける細胞内生化学反応のモデル



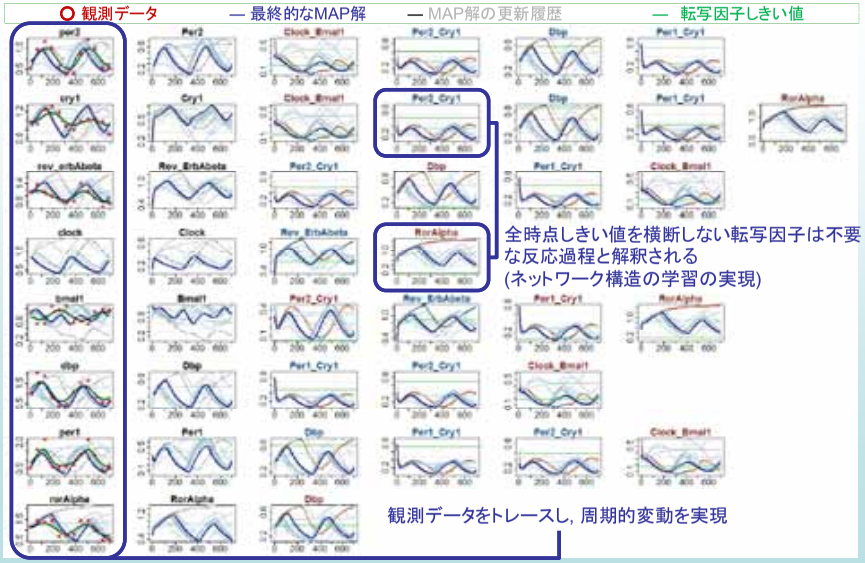
III. 現時点でのソフト研究開発成果

テスト問題: 哺乳動物の概日周期転写ネットワーク

- モデルの規模
  - システム変数の個数: 29
  - 観測変数の個数: 7
  - モデルパラメータ数: 116

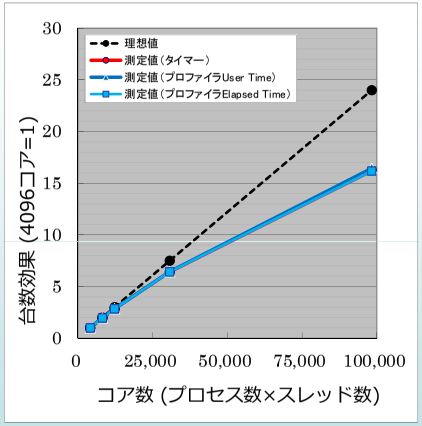


推定パラメータに対するシミュレーション結果

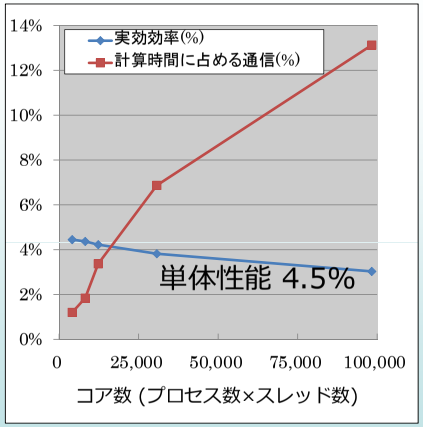


「京」における計算性能

台数効果



実効効率



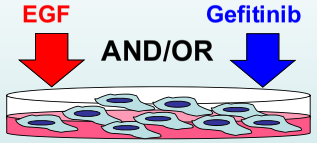
10万並列までの台数効果の伸びを実現した。他方,通信による鈍化が見られるためさらなる性能向上に向けてアルゴリズムを改良中。

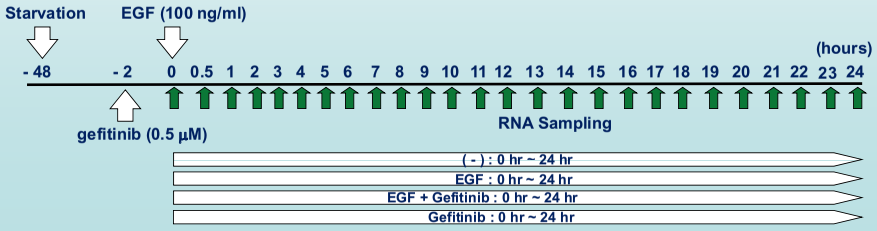
(註:京は現在開発中であり、これらの数字は現状の値である)

10

適用例: 肺がん細胞の薬剤応答

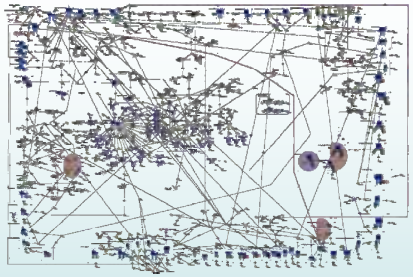
Human Genome Center
Institute of Medical Science, University of Tokyo





演繹的モデル設計の困難と解決策

- 当初は、既知の生化学反応過程を積み重ねて作成したネットワーク・モデルを利用することを試みた。
- しかし、このモデルにはフィードバックが欠落しているため、そのままでは力学システムを構成しないことが明らかとなった。
- TRANSPATHなどのDBを参照して過程を補うことも検討したが、フィードバックを構成するには、既知の過程だけでは不十分であることも分かった。

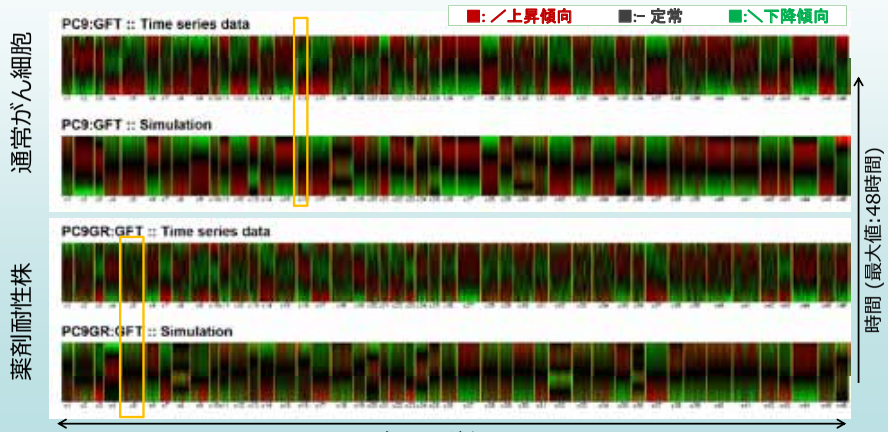


EGF(上皮成長 因子)受容体へのシグナルが細胞内に伝達される過程を記述したモデル (提供: 東大医科研宮野研究室)

- 得られた全17,654遺伝子の発現プロファイルを観察すると、パターンが類似したクラスタに分類できることがわかる。そこで、クラスタに仮想的な遺伝子を割り当て、仮想遺伝子間の制御関係を学習することで、ネットワーク・モデルを構成することを試みた。

シミュレーション結果

- モデルの規模: シミュレーション変数46個, 観測変数 17,654個 (詳細はポスターで)
- 大局的には、観測データをよく再現する。一方、□部分のように、観測に反する遺伝子群も残されている。



データ同化ライブラリ適用事例:
線虫初期胚における細胞内流動データ同化
(遺伝研 木村研究室(細胞建築学)との共同研究)

prior $p(\theta)$ → likelihood $p(y_{1:T} \mid \theta)$ → posterior $p(\theta \mid y_{1:T})$

Model parameters
- velocity of the cell wall
  how to design $u_{wall}(L,t)$
- viscosity coefficient
- density

EnKF
- search model parameter maximizing the likelihood
- observed velocity field

MCMC
- obtain posterior distribution (multi peaked pattern?)

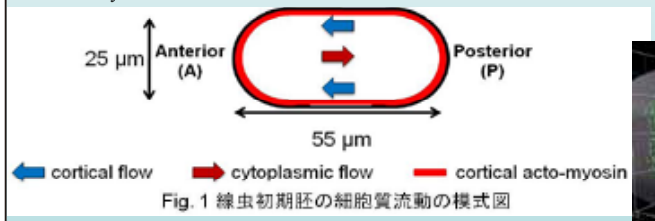
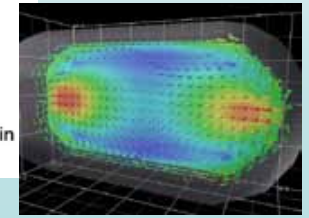


Fig. 1 線虫初期胚の細胞質流動の模式図



IV. プロジェクト終了時のソフト研究開発の達成目標

- 肺がん細胞の薬剤応答を適用例としてLiSDASの開発を進めてきた。しかし、現状のモデルでは、一部の遺伝子群について発現パターンを再現しない。生物学的な知識を追加することで、改良を進める。
- モデルの改良にともなうモデル規模の拡大が予想される。モデル規模が拡大した場合の対応(性能的な意味で)を行う。
  - 最終的なパラメータ数: 368 → 500
  - 内生変数: 17654 (=遺伝子の総数) → 17654 + α
- 10,000ノードを越えると、通信負荷による並列性能の低下が目立ってくる。アルゴリズムを見直し、30,000ノードまで線形に近い台数効果の実現することを目指す。
- 肺がん細胞の薬剤応答モデルの構成法は一般に利用できるものである。現時点の実装では、LiSDASの外部にあるが、これを本体に組み込むことで利便性を向上させる。

謝辞
京での計算に関しては京速コンピュータ京の試験利用、および本年3月 での特別運用での結果です。また、PCクラスタでの性能計測に関しては理化学研究所情報基盤センターのRICCを 使用しています。

啓発活動

(2011年4月刊行)

(2011年9月刊行)

(2011年11月刊行)

企画・総説執筆

データ同化研究開発センターの紹介ビデオ(10分)を作成しました。YouTubeにアップしています。

（謝辞）
本資料集に記載されている「京」での計算は、2011 年 3 月の「京」の特別運用およびその後の試験利用によって行われたものです。

また、本資料集に記載されている「京」を使用した測定値は、開発整備中の「京」による、測定時点での数値です。