H19年12月25日

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

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

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

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

宮野教授PPTから改変

データ解析融合チーム研究開発マップ

（午後に宮野教授から詳しく紹介）

「一般」のデータ

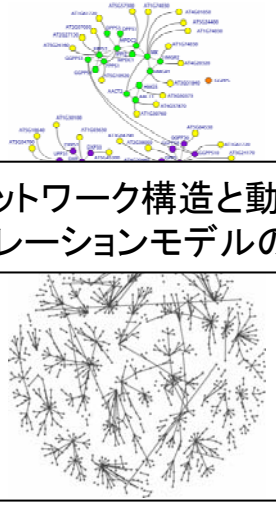
トランスクリプトーム、プロテオームデータをはじめとする生命システムの観測データ。

「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。

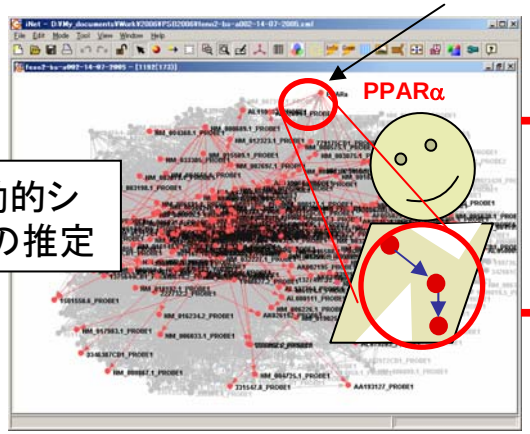
ペタフロップス級の計算によって、創薬ターゲット探索や個人差を考慮した医療開発に貢献する。

「一般」のモデル

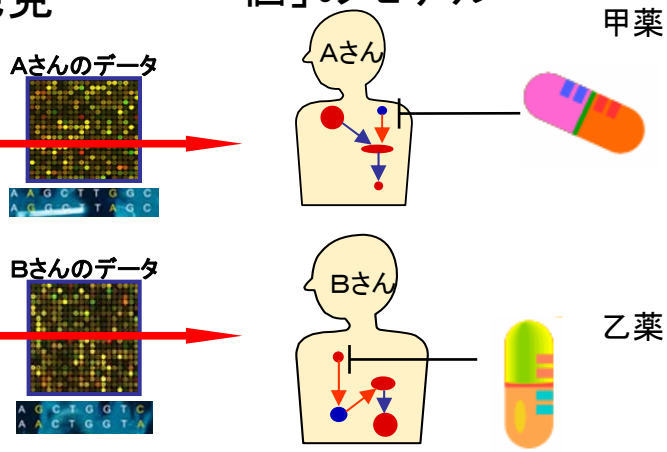


ネットワーク構造と動的シミュレーションモデルの推定

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。

宮野教授PPTから改変

次世代生命体統合シミュレーションの研究開発シンポジウム2007 「セッション①階層を越えるモデリングへの挑戦」

H19年12月25日

データ解析融合チーム

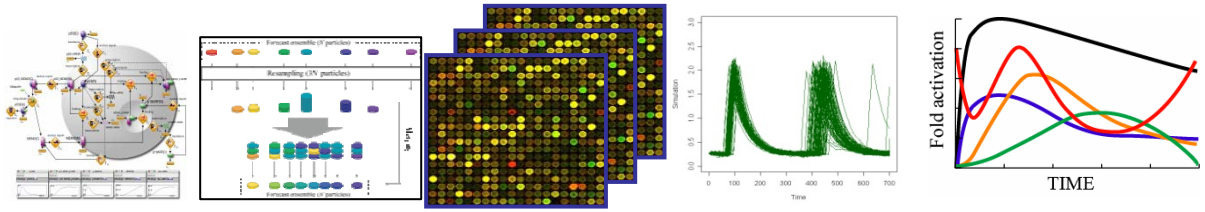
生命体シミュレーションのためのデータ同化技術の開発 (樋口担当)

情報・システム研究機構　統計数理研究所

予測発見戦略研究センター　データ同化グループ

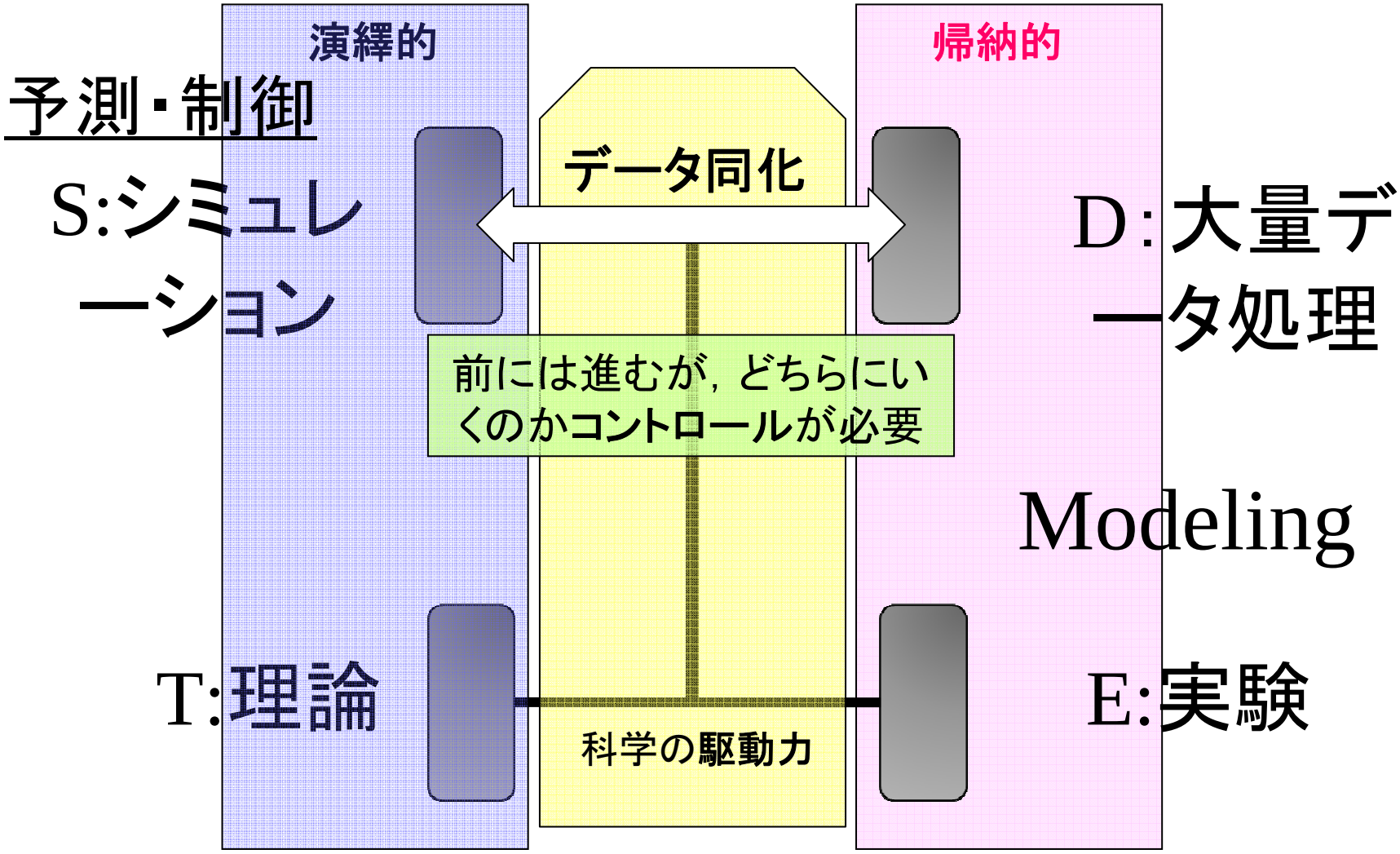
樋口知之，佐藤整尚

上野玄太，吉田　亮



ISM

TESD: 第4の科学，第4の方法論



ISM

データ同化の目的：気象・海洋学の観点から

[1] 予報を行うための最適な初期条件を求める。これは既に、現業の天気予報で実用化されていることである。

[2] シミュレーションモデルを構成する際の最適な境界条件を求める。連成現象を取り扱う際の適応的な境界条件設定もこの作業に含まれる。

[3] スケールが異なるシミュレーションモデル間の橋渡しを行うスキーム内に含まれる諸パラメータの最適な値を求める。経験的に与えられるモデル内のパラメータ値の検証も一つの具体例である。

[4] シミュレーション（物理）モデルにもとづいた、観測されていない時間・空間点における観測値の補間を行う。この作業は再解析データセットの生成とも呼ばれる。このデータセットから新しい科学的発見をもくろむ。

[5] 時間・経費を節約できる効率的な観測システムを構築するための仮想観測ネットワークシミュレーション実験や感度解析を行う。

データ同化を統計科学の視点からとらえる

[1] 予報を行うための最適な初期条件構築
[2] 最適な境界条件設定
[3] パラメータの最適な値の探索と検証
[4] 観測値の補間
[5] 仮想観測ネットワークシミュレーション実験や感度解析

これらは統計科学の手法ーアルゴリズムやモデリングー開発の研究課題

(A) フィルタリング（予測手法を含む）: [1], [2] が主に対応

(B) 最適化: [1], [2], [3]が対応。特に [3]

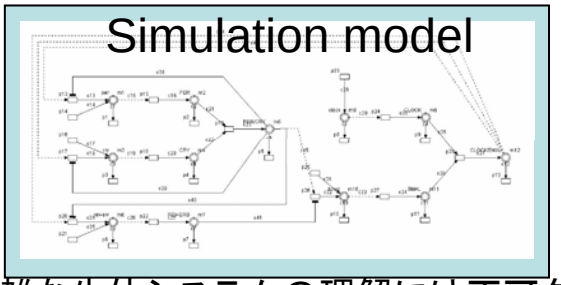
(C) 内挿・外挿（平滑化アルゴリズム）: [4]

(D) 計測（観測）システムのデザイン（誤差評価）: [5]

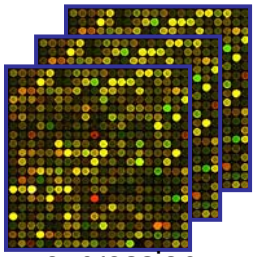
(E) 知識発見: [4]

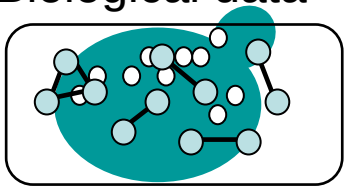
生命体データ同化技術

情報統合のプラットフォーム：シミュレーションモデルとデータが直接会話する場



＋





•複雑な生体システムの理解には不可欠
•どんなに詳細にしても不完全
–現実との乖離、個別ケースの予測能力低下

・超ヘテロ，超高次元，
含ノイズ，欠測，部分的

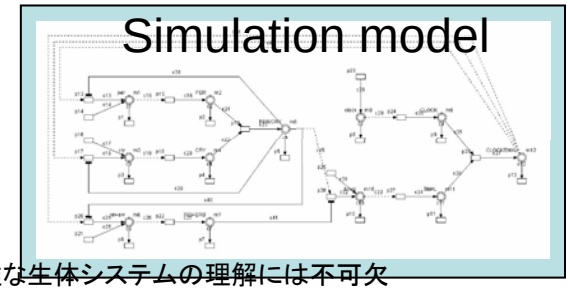
[データ同化による解決]

- シミュレーションと現実とのギャップを埋める枠組み
- シミュレーション結果と観測データを、按配よくブレンド
  - 統計的手法（ベイズ統計）
    ○モデル・データ双方の不確実性を反映
  - 現実への接近　予測能力向上
  - パラメータの自動推定・モデル選択
- モデル・データ双方単独では抽出できない情報を抽出

ISM

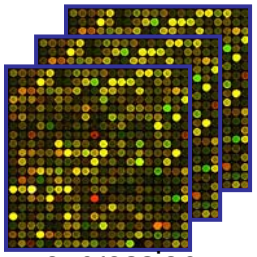
シミュレーションを統計科学の枠組みに埋め込む

情報統合のプラットフォーム：シミュレーションモデルとデータが直接会話する場

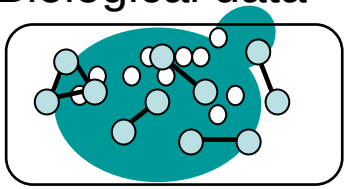


•複雑な生体システムの理解には不可欠
•どんなに詳細にしても不完全
–現実との乖離、個別ケースの予測能力低下

＋



Biological data



・超ヘテロ，超高次元，
含ノイズ，欠測，部分的

一般状態空間モデルによる定式化

$$\mathbf{m}_t = f(\mathbf{m}_{t-1}, \mathbf{w}_t, \boldsymbol{\theta})$$ シミュレーションモデル

$$\mathbf{y}_t = H\mathbf{m}_t + \boldsymbol{\varepsilon}_t$$ 観測モデル

$$t = 1, \ldots, T$$

$\mathbf{m}_t$：時刻 t での状態ベクトル　f：シミュレーションコードが規定する時間発展

$\mathbf{w}_t$：システムノイズ　$\boldsymbol{\theta}$：パラメータベクトル

$\mathbf{y}_t$：時刻 t での観測ベクトル　H：観測行列

$\boldsymbol{\varepsilon}_t$：観測誤差

ISM

パーソナライゼーション技術への対応を視野にいれて

「パーソナライゼーション」ニーズの高まりの背景：

- 資源の有効利用のために選択と集中
- 価値観の多様化
- “コ”（個人，個性，固有，個別）に特化

大量生産・大量消費をめざした20世紀→
個人に焦点をあわせる科学へ

■オーダーメイド医療，副作用の研究，マイクロマーケティング，One-to-One　＊，Situation　＊，環境に優しい商品

“個”にマッチしたシミュレーション：
境界条件の設定機能をパーソナライズする

“個”によって異なる形状，形態情報をシミュレーション
モデルに取り込む 『メタシミュレーションモデル』

運動方程式：$\mathbf{v}$

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v}\cdot\nabla)\mathbf{v} + \mathbf{f}\times\mathbf{v} = -g\nabla\eta - \gamma_b \frac{\mathbf{v}|\mathbf{v}|}{H} + A_H\nabla^2\mathbf{v}$$

海底摩擦

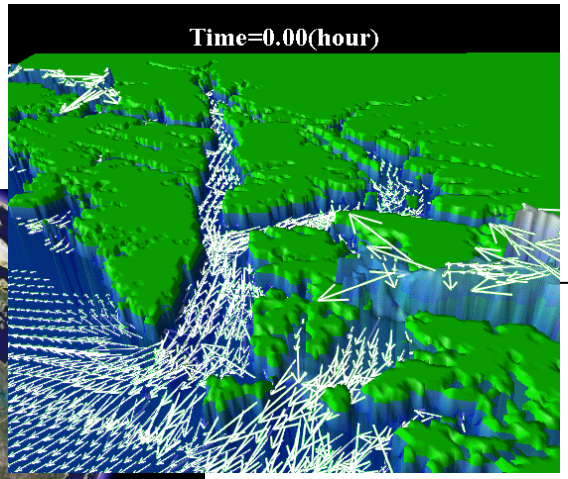
海底摩擦係数

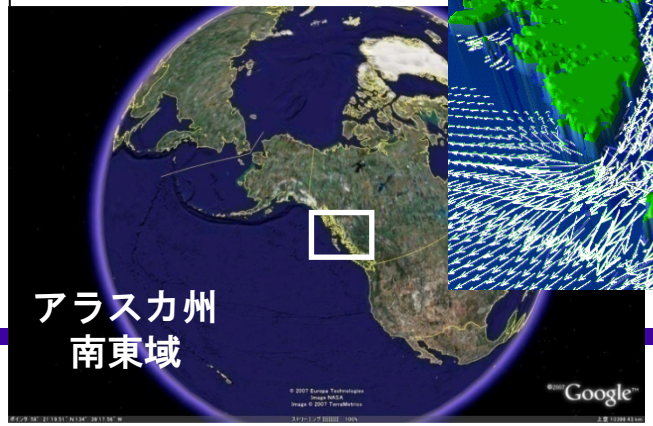
水深

連続式：

$$\frac{\partial \eta}{\partial t} + \frac{\partial}{\partial x}(\mathbf{v}H) = 0$$

最適摩擦係数
γ_b : 0.006





$\mathbf{v}$： 水平（2次元）流速ベクトル
η： 海面水位
H： 水深，$\mathbf{f}$： コリオリパラメータ

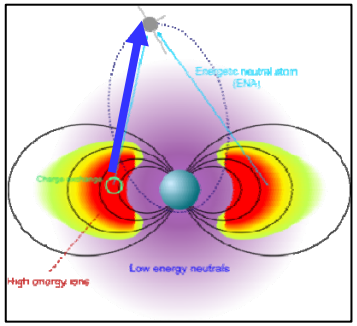
我々研究チームによる，潮汐シミュレーションの例

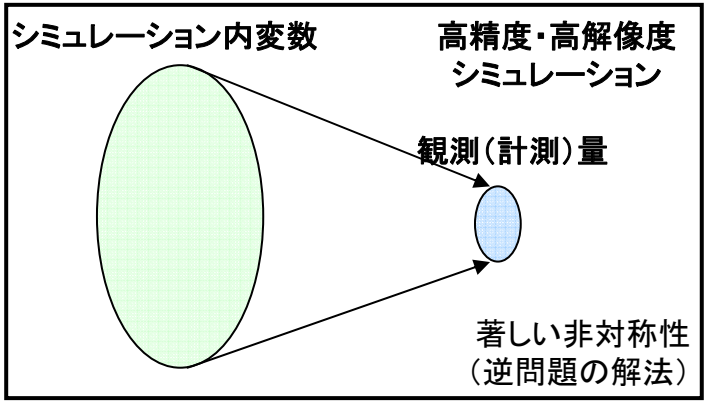
超高精度医療動画像技術と共進化するシミュレーション技術開発をめざして

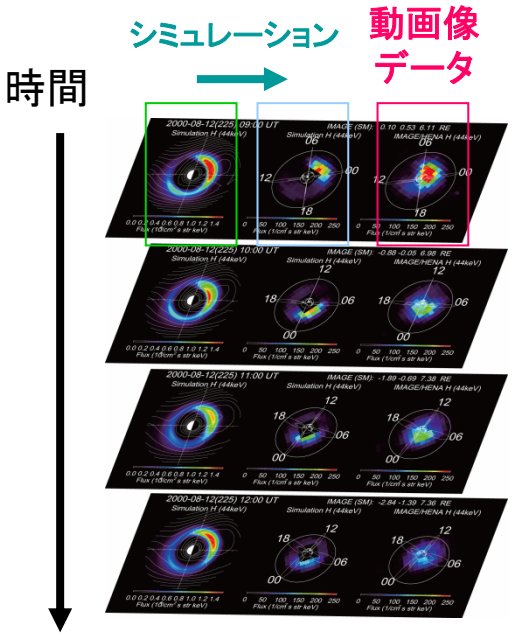
・ハイブリッドシミュレータ（計測情報を直接適切に取り込む）

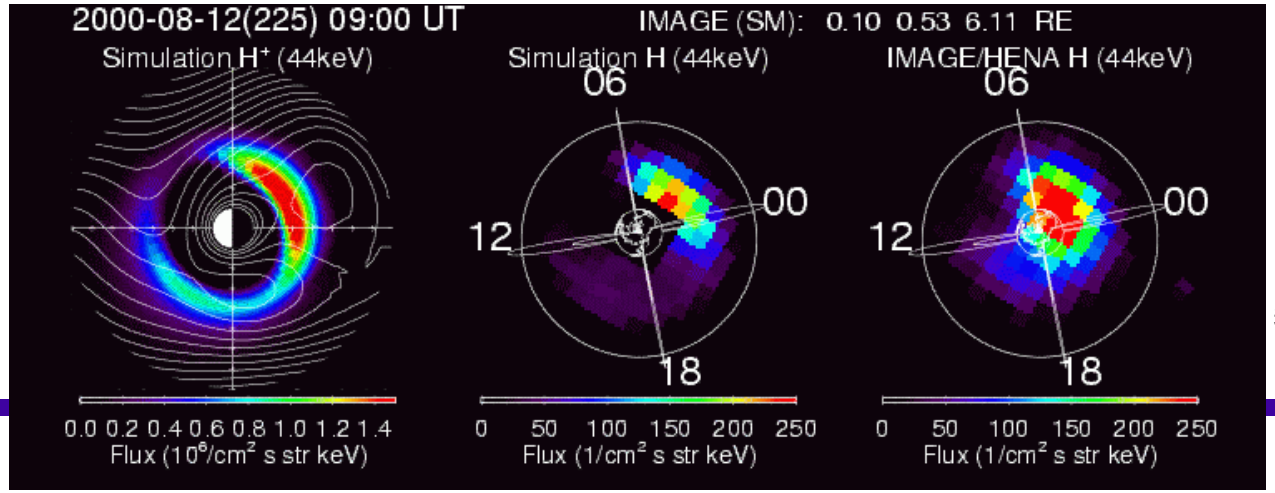
・直接観測できない対象の非侵襲的方法による計測データ

・元のシミュレーションモデルではうまく表現できなかった時間変化を再現









我々研究チームによる，世界最大規模(200万次元)の逐次データ同化手法の成功例

ISM

『生命体データ同化技術』の開発

・どんなプログラムか

－シミュレーションとデータ解析を統合

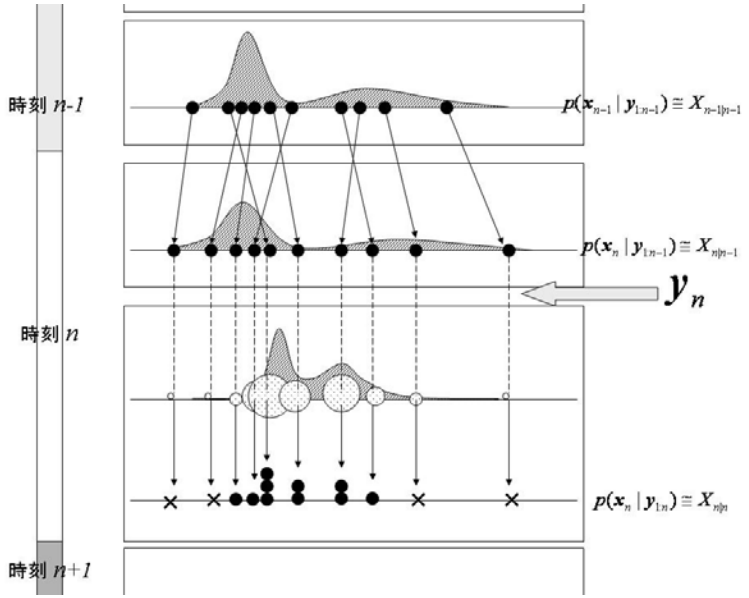
－予測，リスク解析を行うためには**分布の見方**（視点）が大切で，そのためには同質の計算を大量に高速に行う必要

－従って，分布を精度良く推定することを主眼とした，計算資源の選択と集中策の開発

・何ができるのか

－データ駆動型イン・シリコネットワークモデリングの自動化

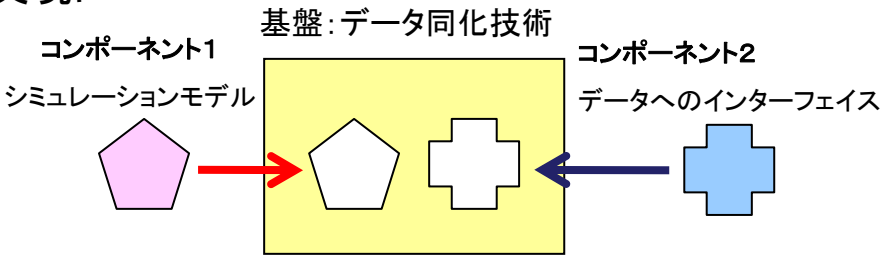
観測データに即したシミュレーション：観測データをペタスケールコンピュータ上で学習させ，シミュレーションモデルのパラメータ探索，仮説ネットワークモデルの生成の自動化が可能に



図：逐次データ同化技術：アンサンブルベースフィルタ.

・将来は

－個人，環境に適した**パーソナライズ**されたシミュレーションにもとづく解析と予測が今後重要．データ同化はその実現に必須の技術．



・誰が使うのか

－開発段階研究者

－完成時にはCOE的な先端的医療病院の技師

ISM

『生命体データ同化技術』の開発(2):循環と発展

・これまでの開発状況と今後の計画

ーH19年: データ同化技術の開発においてベースとなるネットワークモデルの調査と選定

ーH20年: 逐次凸最適化法と粒子フィルタによるパラメータ推定技術の開発

ーH21年: 生体内ネットワークの統計的推論における自動仮説生成システム(統計的モデル選択法)の構築

ーH22年以降: 大規模生体内ネットワーク推定を可能にするデータ同化技術の開発・強化

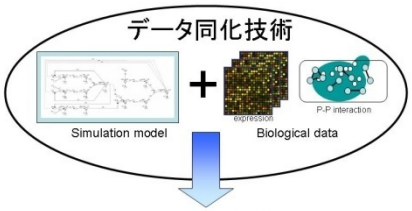
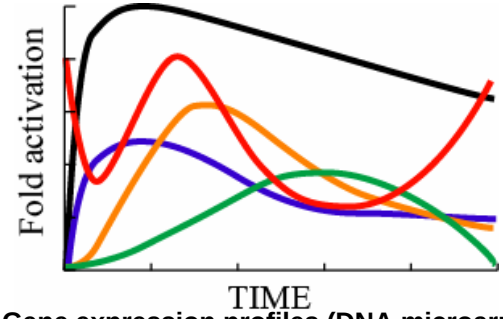


図:システム創薬のための基盤技術開発

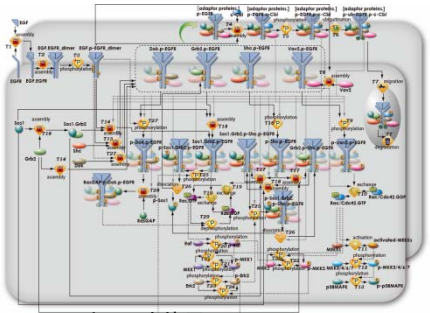


- **Gene expression profiles (DNA microarray)**
- **Proteomic data (mass spectrometry)**

- 仮説の生成
- 実験計画

- 力学的パラメータの推定
- モデルの評価と選択
- リモデリング

Simulation models



・現状での計算規模と次世代スパコンでの計算規模

ー現状:200万次元, 粒子数1500個, データ800次元, データ数100ポイント で**1日**

HP ProLiant DL145 G2(AMD Opteron 2.6GHz 256CPU 640G 1.3TFLOPS, (64CPU, 276GB)のキューを使用

ー次世代スパコンでは, **数秒**で計算. → **個人化**に対応可能

ISM