

ライフ・テクノロジー

～生体ビッグデータの解析を通して、健康
安心・安全な生活を実現する～

産業技術総合研究所

生体システムビッグデータ解析オープンイノベーションラボラトリ

副ラボ長

油谷 幸代

研究紹介

- ・ 目標 :

世界標準となる新規生命情報解析技術の開発

情報技術による生体疾病メカニズム解明

情報技術による創薬・有用化合物シース探索と微生物生産制御の実現

情報技術による生体行動メカニズム・個体健康差の要因解明

- ・ 社会実装イメージ :

第3世代NGSデータの標準情報解析ツールの提供 → 医薬産業へ

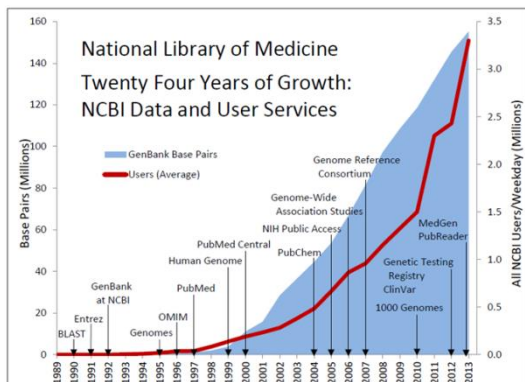
スマートセル情報解析プラットフォーム構築 → 化成品・食品製造産業へ

免疫・生活習慣病に効能が高い高機能製剤開発のための情報提供 → 健康・医薬業界へ

生体メカニズム解明による健康維持のための情報提供 → 健康・医薬産業へ

生命系データ量の推移

NCBIゲノム情報量推移



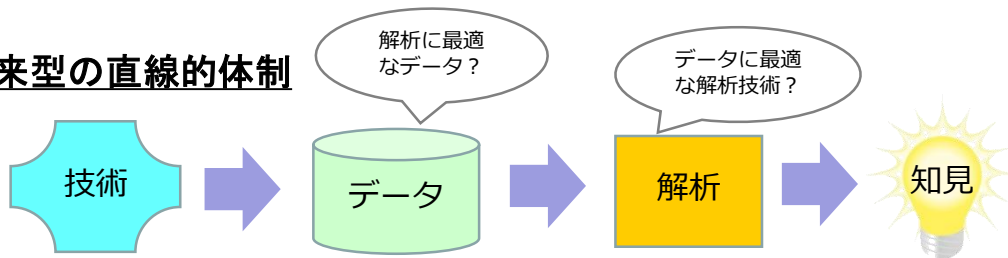
<http://www.nlm.nih.gov/about/2015CJ.html>

実験技術の進展により、

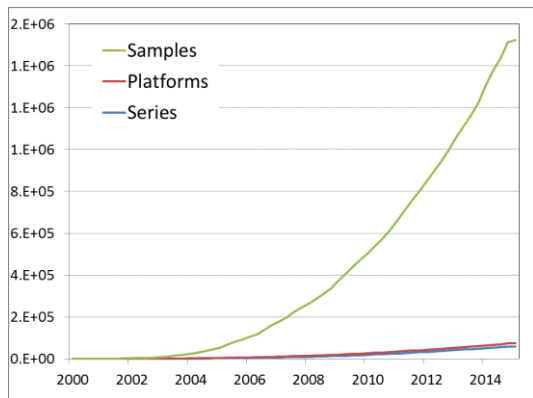
今後もデータ量は指数関数的に増加の予測

⇒効率的に知見を得るためには情報学的手法が必須

従来型の直線的体制

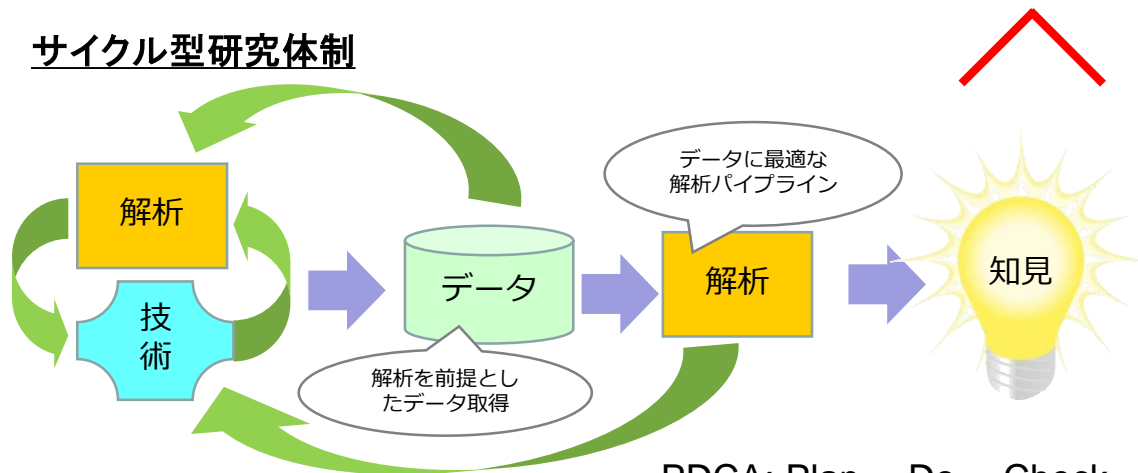


NCBI遺伝子発現情報量推移



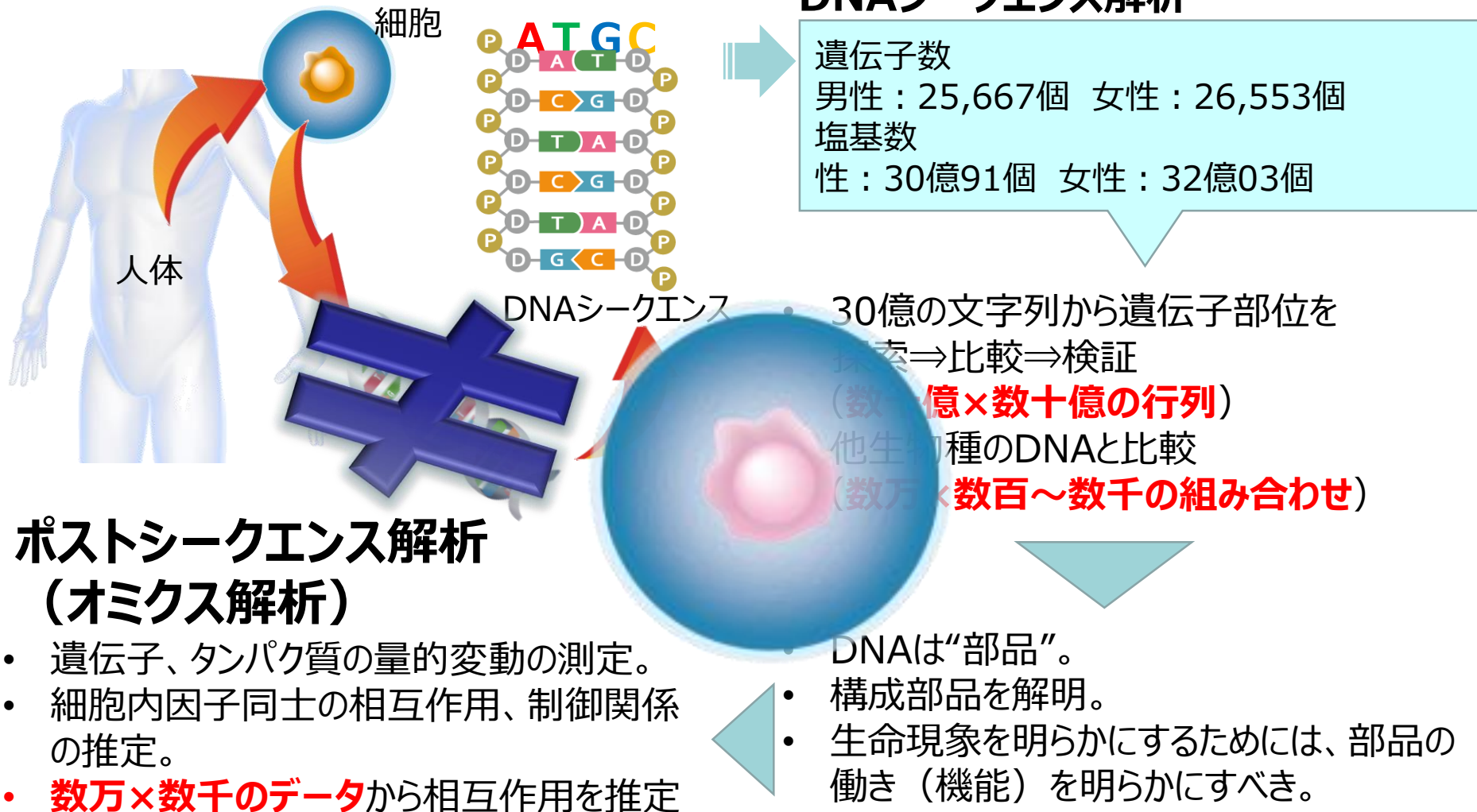
<http://www.ncbi.nlm.nih.gov/geo/summary/?type=history>
より作成

サイクル型研究体制



PDCA: Plan→ Do→ Check→ Act

生体システムデータの歴史



統合オミクスから生体システム解明へ

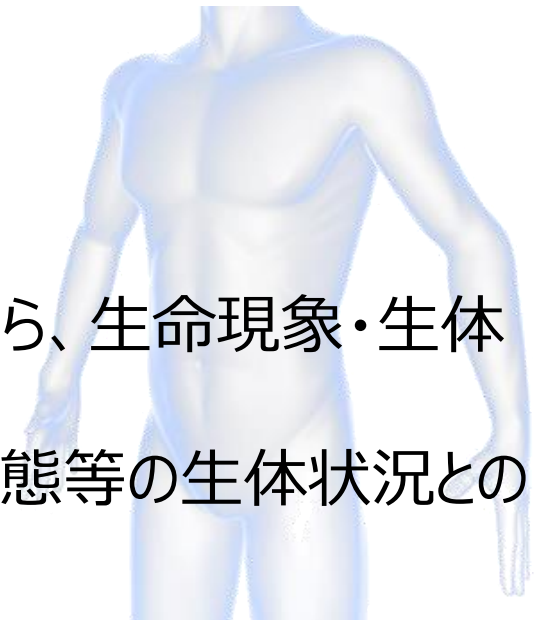
統合オミクス解析

- 遺伝子とタンパク質、代謝産物など**異なった階層間での相互作用**を推定。
- 疾病や健康状態等の生体状況と各種オミクス情報を統合的に解析。



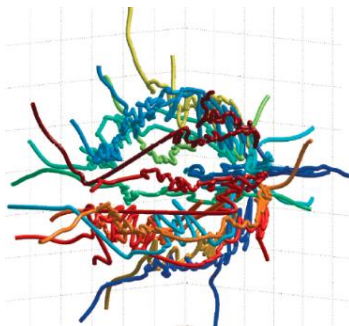
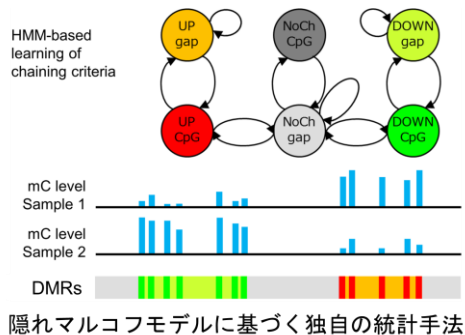
生体システムの解明

- システムティックに測定された様々なデータから、生命現象・生体状況のメカニズムを解明。
- 様々な生体情報を観測し、疾病や健康状態等の生体状況との相関を推定。



配列情報解析

- **AIを使ったゲノムビッグデータ解析技術の開発**
 - アラインメントパラメタの自動学習 (LAST-TRAIN)
- **エピゲノムビッグデータ解析技術の開発**
 - DNAメチル化変化の検出
 - ゲノム高次構造変化の検出
- **非コードRNAビッグデータ解析技術の開発**
 - 超高速RNA-RNA相互作用予測 “(RIblast)
- **生体ビッグデータの安全な解析技術の開発**
 - 医療テキストの安全な検索



Start with a rough guess of the probabilities

Sequences

```
cgtagctgctatgcgtagctag
cgtatgctgtaagatgtagtcg
tttgatgatatctctattg
cgcgcgcgctagcgcgcc
```

LAST-TRAIN

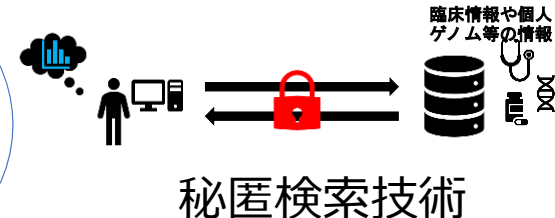
Substitution & gap probabilities

```
a c g t
a .22 .016 .044 .018
c .016 .13 .011 .044
g .044 .011 .13 .016
t .018 .044 .016 .22
gap existence probability: 0.015
gap extension probability: 0.73
```

Alignments

```

- - - - -
| | | | |
a c a t g c t a g g
```

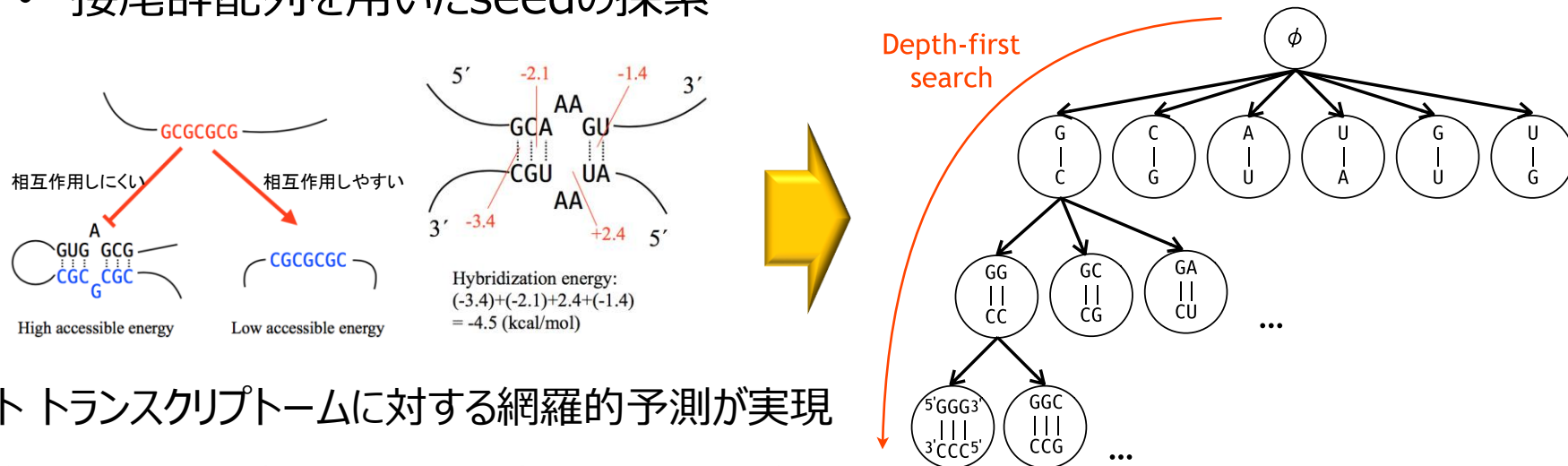


配列情報を医薬分野へ：lncRNAと疾病メカニズム

Riblast 超高速RNA-RNA相互作用予測

Fukunaga & Hamada, Bioinformatics, JCB (2017,2018)

- **構造を考慮**した相互作用エネルギーの最小化
- BLASTと同様のseed-and-extend方式を採用（p-valueの計算も可能）
- 接尾辞配列を用いたseedの探索



ヒト トランスクリプトームに対する網羅的予測が実現

計算時間	配列アラインメント	RNA相互作用予測
遅い	Ssearch	IntaRNA RNAplex
seed-and-extendによる高速化	BLAST	Riblast

配列情報を医薬分野へ：lncRNAと疾病メカニズム

LncRRIssearch <http://rtools.cbrc.jp/LncRRIssearch/> lncRNA-RNA相互作用DBの構築

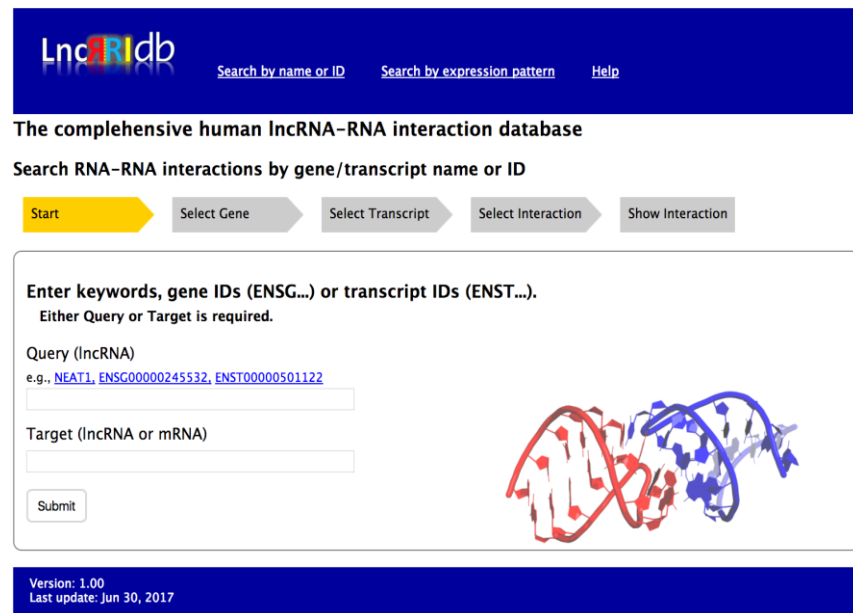
疾患関連lncRNA/mRNAの相互作用を明らかにすることで、疾病メカニズムの解明や創薬につながる！

- 独自開発の予測ツール**RIblast**で推定したRIblastにより予測した「ヒト」と「マウス」のlncRNA-RNA相互作用を格納

→ 唯一の相互作用データ

- 発現データ・局在データと合わせて検索可能
 - － 発現データ：特定の組織で特異的に発現上昇／減少するlncRNA-RNA相互作用
 - － 局在：共に核（あるいは細胞質）に局在するlncRNA-RNA相互作用

→ 生体レベルでの疾病メカニズムへ



LncRRIdb [Search by name or ID](#) [Search by expression pattern](#) [Help](#)

The comprehensive human lncRNA-RNA interaction database

Search RNA-RNA interactions by gene/transcript name or ID

Start Select Gene Select Transcript Select Interaction Show Interaction

Enter keywords, gene IDs (ENSG...) or transcript IDs (ENST...).

Either Query or Target is required.

Query (lncRNA)

e.g., [NEAT1](#), [ENSG00000245532](#), [ENST00000501122](#)

Target (lncRNA or mRNA)

Submit

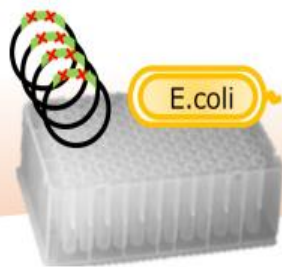
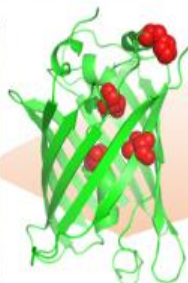
Version: 1.00
Last update: Jun 30, 2017

<http://rtools.cbrc.jp/LncRRIssearch/>

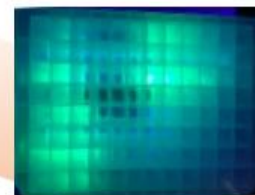
配列情報を医薬分野へ：人工知能によるタンパク質の機能改良

- バイオ医薬品、酵素など様々な機能性タンパク質の開発を加速

緑色蛍光タンパク質



学習用ライブラリー



機能評価

遺伝子配列
蛍光機能
発現量

教師データ
155 個

黄色蛍光タンパク質



機械学習
ベイズ最適化

機械学習提案ライブラリー

報道： 日本経済新聞電子版、化学工業日報、日刊工業新聞
産総研プレスリリース (2018 / 8 / 31)

配列情報を医薬分野へ：次世代配列解析技術の開発

● 疾患関連タンデムリピートの発見

- S Mitsuhashi, MC Frith, et al. [Genome Biology](#) 2019
- J Sone, S Mitsuhashi, ..., MC Frith, N Matsumoto, G Sobue. [Nature Genetics](#) 2019

● 長鎖DNAからのリアレンジメントの発見

- MC Frith, S Khan. [Nucleic Acids Research](#) 2018

● 配列アラインメントの基礎理論

- MC Frith. [Bioinformatics](#) 2019

● ナノポアシーケンサーのデータ解析技術開発

- S Mitsuhashi, ..., MC Frith, H Mitsuhashi. [Scientific Reports](#) 2017
- M Seki, ... MC Frith, K Tsuchihara, Y Suzuki. [DNA Research](#) 2019



新テクノロジーのための情報解析技術開発と医療分野への応用

リアルデータを医療応用へ

● 医薬系リアルデータの情報解析

－ メタゲノム解析

- **腸内細菌叢** 巨大混合ゲノムデータの解析
- 細胞内外の共生系

－ 疾病発生・がん転移メカニズムの解析

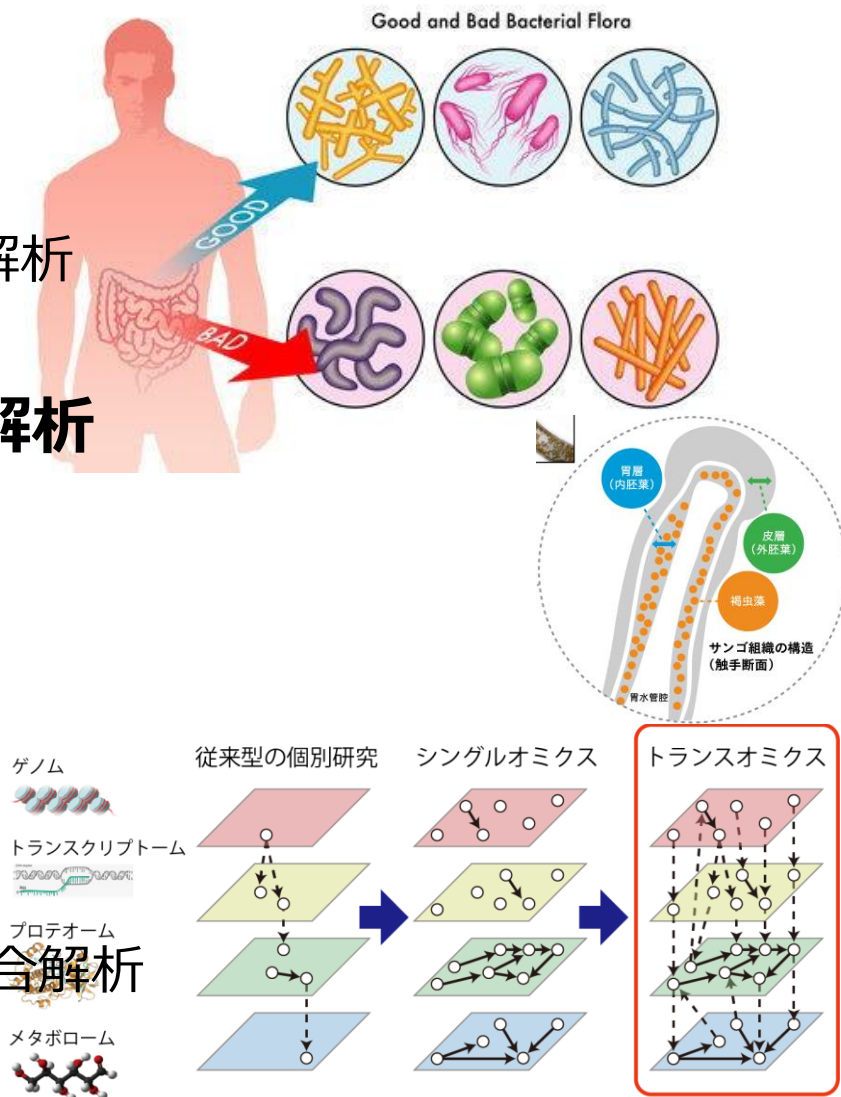
- 空間動態の解析
- 転移マーカーの同定

－ シングルセルの情報解析

- 難培養細菌解析
- **腸内細菌叢**、高精度アセンブリ技術

－ マルチオミクス解析

- 巨大遺伝子発現データと代謝系の統合解析
- 微生物による有用物質生産の最適化

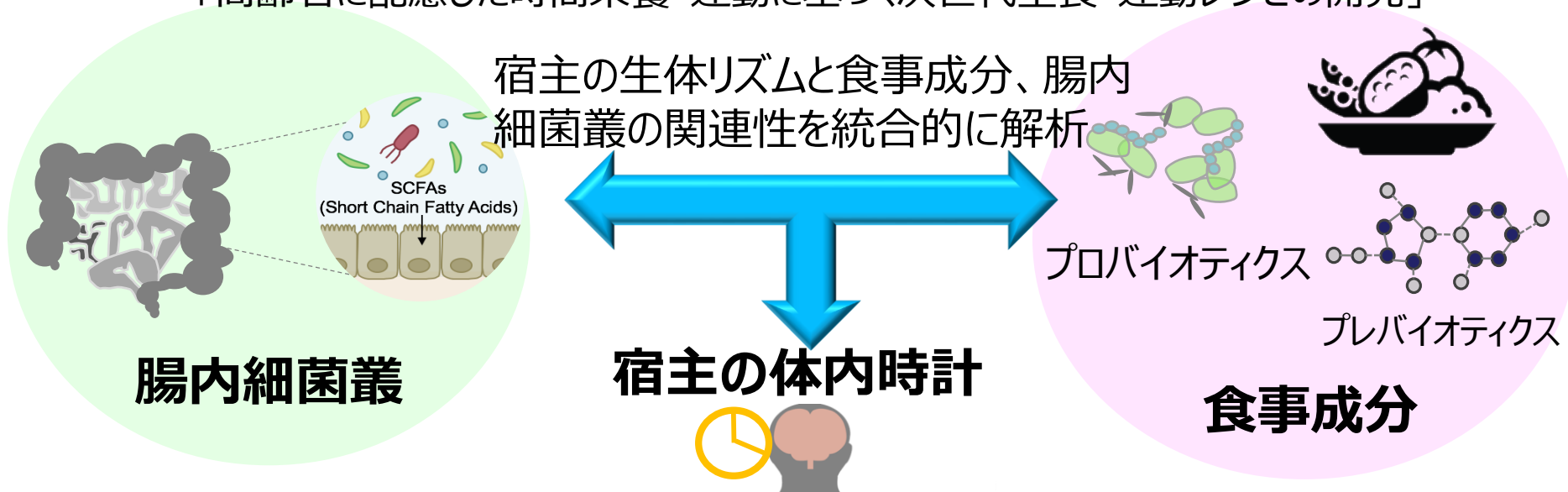


リアルデータを医療応用へ：腸内細菌叢を医薬品へ

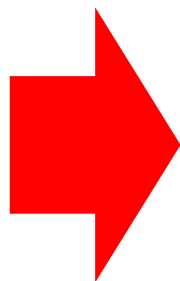
・時間栄養学に基づいた生体内腸内細菌叢解析

戦略的イノベーション創造プログラム(SIP)採択事業

「高齢者に配慮した時間栄養・運動に基づく次世代型食・運動レシピの開発」



・免疫
・生活習慣病
・肥満
との関連を推定



・生体レベルのコントロールシステムが
明らかに
・**効能が高い高機能製剤開発へ**

リアルデータを医療応用へ：先端技術データ解析

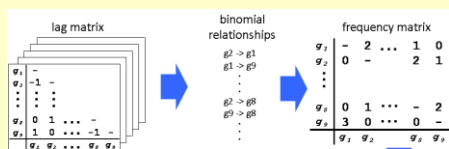


早大生命医薬系
研究室群

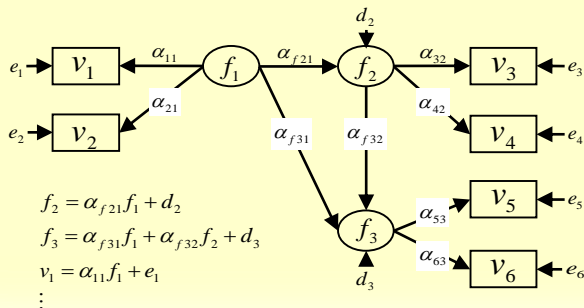
微量組織からの遺伝子発現・位置情報データ



細胞間相互作用推定技術の開発

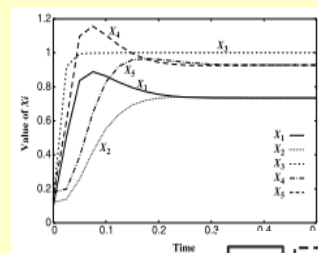


$$\Sigma = GT\Sigma_u T'G'$$

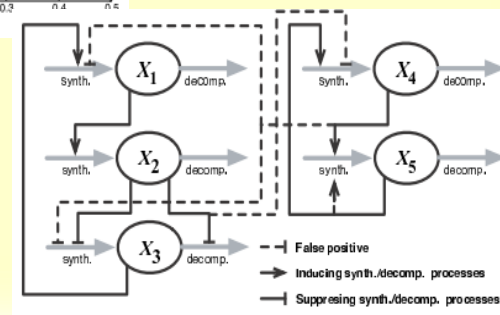


CBBB-OIL

細胞間相互作用の動的解析
(ダイナミクス)



$$\frac{dX_i}{dt} = \alpha_i \prod_{j=1}^n X_j^{g_{ij}} - \beta_i \prod_{j=1}^n X_j^{h_{ij}}$$



組織への影響メカニズム解明

細胞から生体挙動の解明へ:生体システム制御

・数値データからのネットワーク解析と生体モデル解析

1. システムの設計図を明らかにする

→ネットワーク構造解析

- ✓ 化学物質の毒性発症メカニズム解析
- ✓ 外部要因と細菌叢変動の因果関係解析
- ✓ 微生物における有用物質生産メカニズム解析



疾病ターゲット同定へ

2. システム制御を明らかにする

→微生物生産モデル構築・ダイナミクス解析

- ✓ 生産制御のためのネットワーク解析
- ✓ 生産制御のための微生物生産モデル構築と解析
- ✓ 代謝流量を計算するダイナミクス解析



医薬品の生産制御へ

生命現象をシステムレベルで理解する

複雑な「システム」を理解し活用するために必要なシステムズバイオロジーストラテジー

1. システムを構成している部品は何か？

1. 遺伝子領域の情報
2. 必要十分に高精度なゲノム配列情報

遺伝子同定
配列決定
関連遺伝子推定

2. 個々の部品がどのような関係を持つか？

1. 有意差解析に資する遺伝子発現プロファイル（RNA-Seq）
2. 変数以上の測定点を有する遺伝子発現データ

ネットワーク構造解析

3. 全体がいつどのように動作しているか？

1. 時系列変動が明らかになっている遺伝子発現データ

動態解析
ダイナミクス

4. 設計し、機能を再現できるか？

配列設計
合成生物学

生物データを数理モデルへ

実験的手法による遺伝子発現状態の測定



測定データの正規化（標準化）

有意差解析

t-検定
Mann-Whitney U検定
ANOVA
SVM

遺伝子群抽出

クラスタリング

階層的クラスタリング
SOM
K-mean

部品から設計図をつくる

ネットワーク構造 動態解析

Boolean Model
Bayesian Model
Graphical Gaussian Model
Graphical Chain Model
Structural Equation Model

ネットワーク解析

動態解析

微分方程式 Model

部品を明らかにする

設計図から制御システムをつくる

構造方程式モデルの必要性

生命現象を数理モデルとして表現するためには…

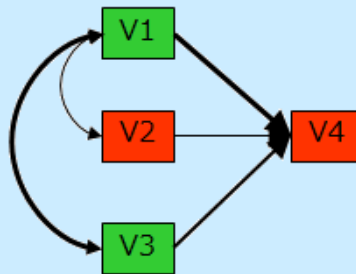
遺伝子だけでなく、
他の細胞内因子を含んだネットワークモデルが必要



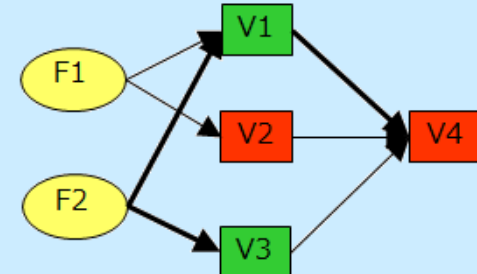
構造方程式モデリングの生命データへの適用

	Exp1	Exp2	ExpP
V1	V_{11}	V_{12}	$\dots V_{1p}$
V2	V_{21}	V_{22}	$\dots V_{2p}$
V3	V_{31}	V_{32}	$\dots V_{3p}$
V4	V_{41}	V_{42}	$\dots V_{4p}$

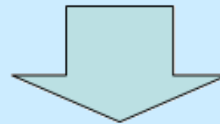
遺伝子発現データ



$V1 \Leftrightarrow V3$: 強い相関
 $V1 \Rightarrow V2$: 弱い相関



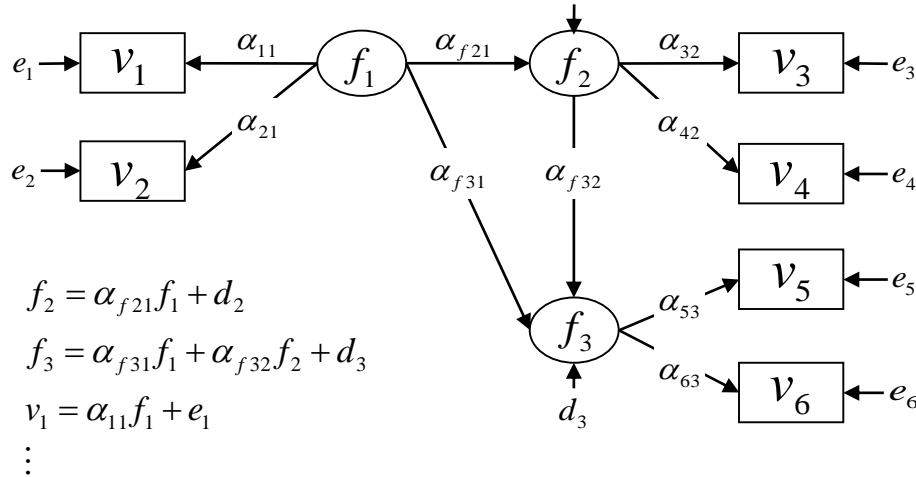
$V1, V2$ の相関の原因 : F1
 $V1, V3$ の強い相関の原因 : F2
 $F1, F2 = \text{タンパク質}$



観測変数間に相関をもたらす、観測されていない要因を潜在変数として仮定し、ネットワークモデル中に観測変数と同一レベルで組み込む

構造方程式モデリング (SEM) 概略

1. 初期モデル



2. 構造式

$$\begin{bmatrix} f_1 \\ \vdots \\ f_3 \\ v_1 \\ \vdots \\ v_6 \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & & & \\ \alpha_{f31} & & 0 & & 0 \\ \alpha_{11} & \cdots & 0 & \cdots & 0 \\ \vdots & & & & \\ 0 & \cdots & \alpha_{63} & \cdots & 0 \end{bmatrix} \begin{bmatrix} f_1 \\ \vdots \\ f_3 \\ v_1 \\ \vdots \\ v_6 \end{bmatrix} + \begin{bmatrix} d_1 \\ \vdots \\ d_3 \\ e_1 \\ \vdots \\ e_6 \end{bmatrix}$$

$$t = At + u \quad t = \begin{bmatrix} f \\ v \end{bmatrix} \quad A = \begin{bmatrix} A_a & A_d \\ A_b & A_c \end{bmatrix} \quad u = \begin{bmatrix} d \\ e \end{bmatrix}$$

5. モデルの分散共分散構造

$$\Sigma_m = GT\Sigma_u T'G'$$

$$\Sigma = \Sigma_m$$

$$= G \begin{bmatrix} I - A_a & -A_d \\ -A_b & I - A_c \end{bmatrix}^{-1} Cov_e \begin{bmatrix} I - A_a & -A_d \\ -A_b & I - A_c \end{bmatrix}^{-1} G^t$$

モデルで定義される分散共分散行列が観測値から計算される分散共分散行列にフィットするようにパラメータ推定を行う

3. 変数の式変換

$$T = (I - A)^{-1}$$

$$t = Tu$$

4. 観測変数の抽出

$$\text{if } G = \begin{bmatrix} O & I \end{bmatrix} \quad v = \begin{bmatrix} O & I \end{bmatrix} \begin{bmatrix} f \\ v \end{bmatrix}$$

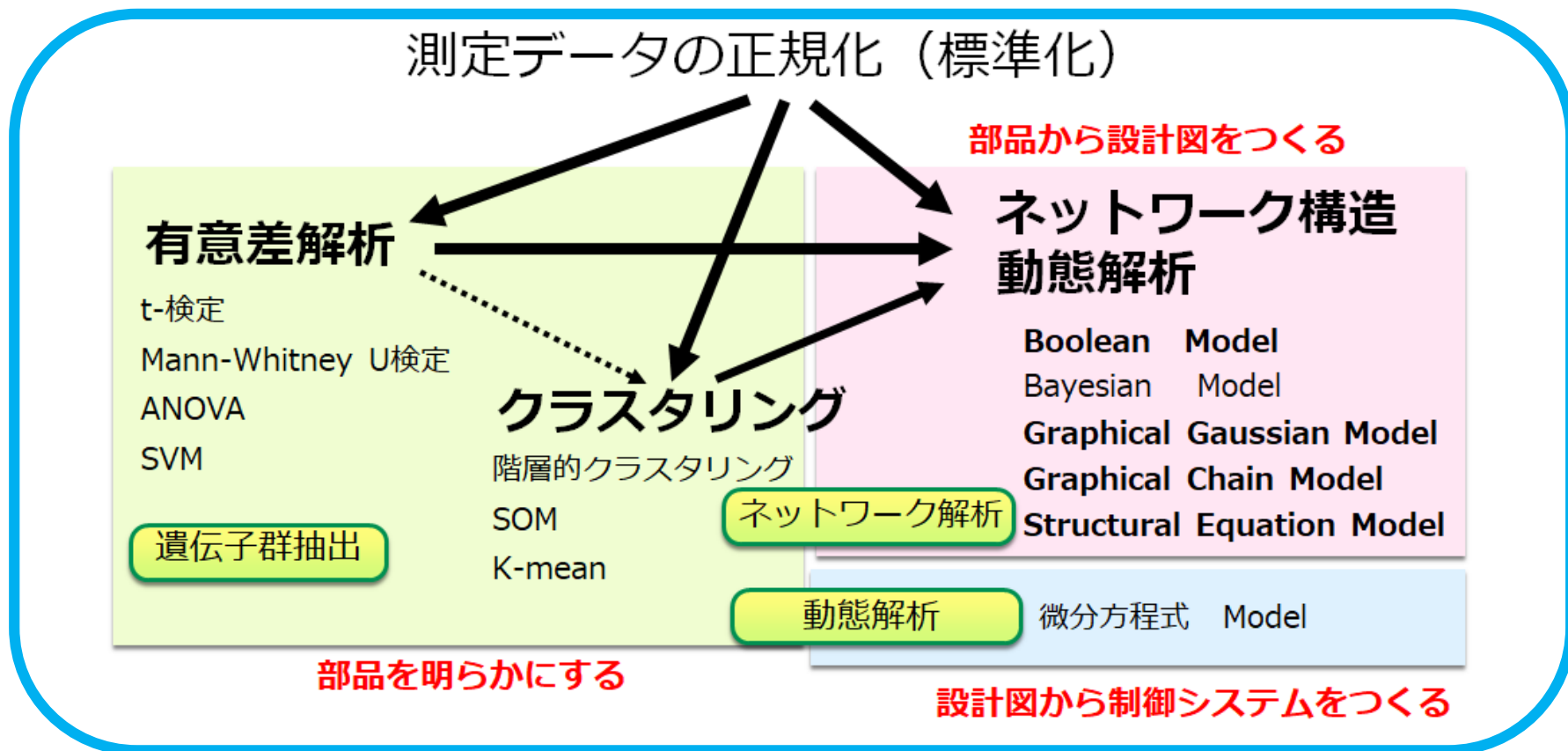
substitute $t = Tu$ for $v = Gt$

$$v = GTu$$

システム制御と設計

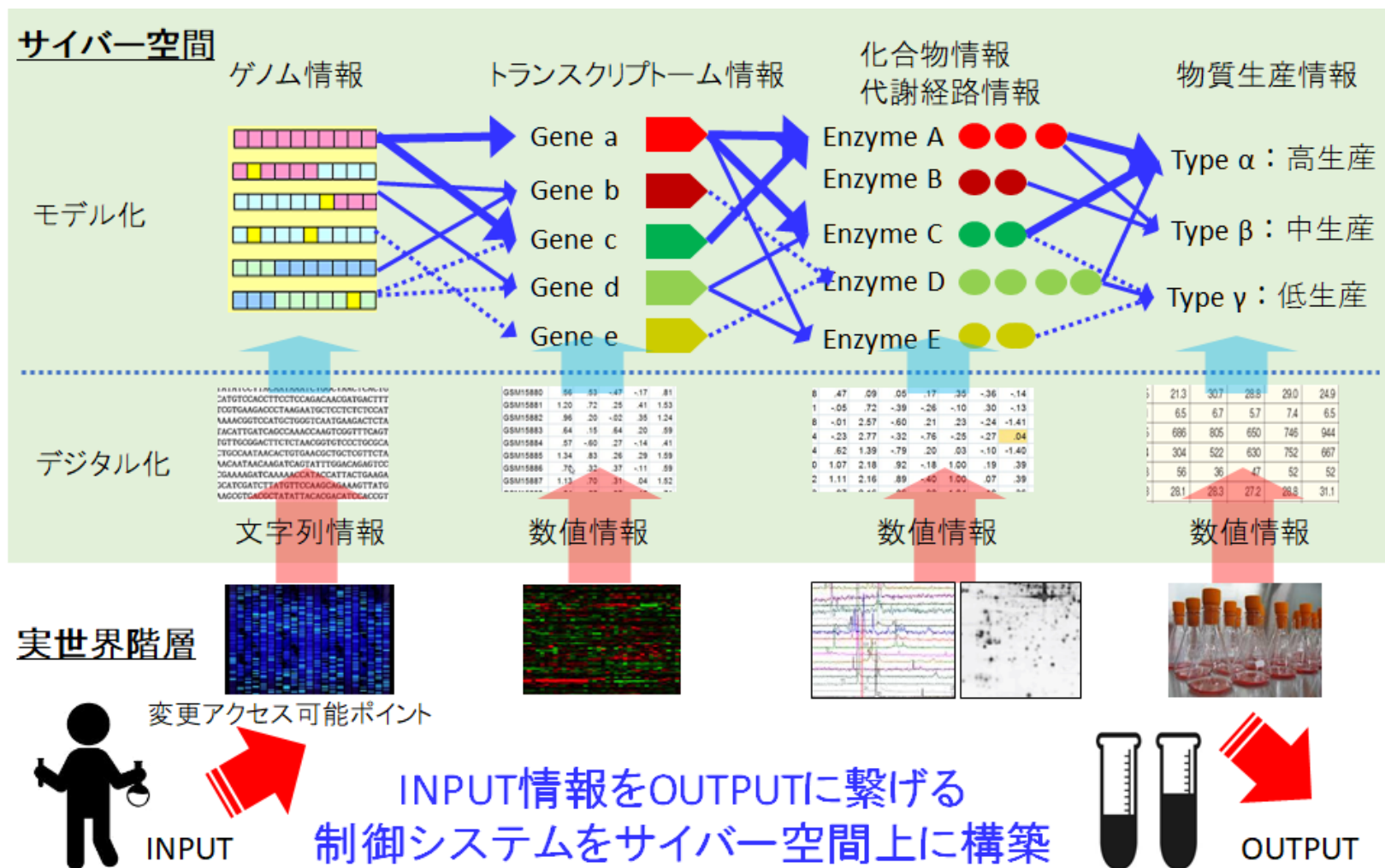
方向性：

1. 複雑な生命現象の操作・改変を目指す→システム制御
2. 複雑な生命現象の動作原理そのものの理解を目指す→システム設計



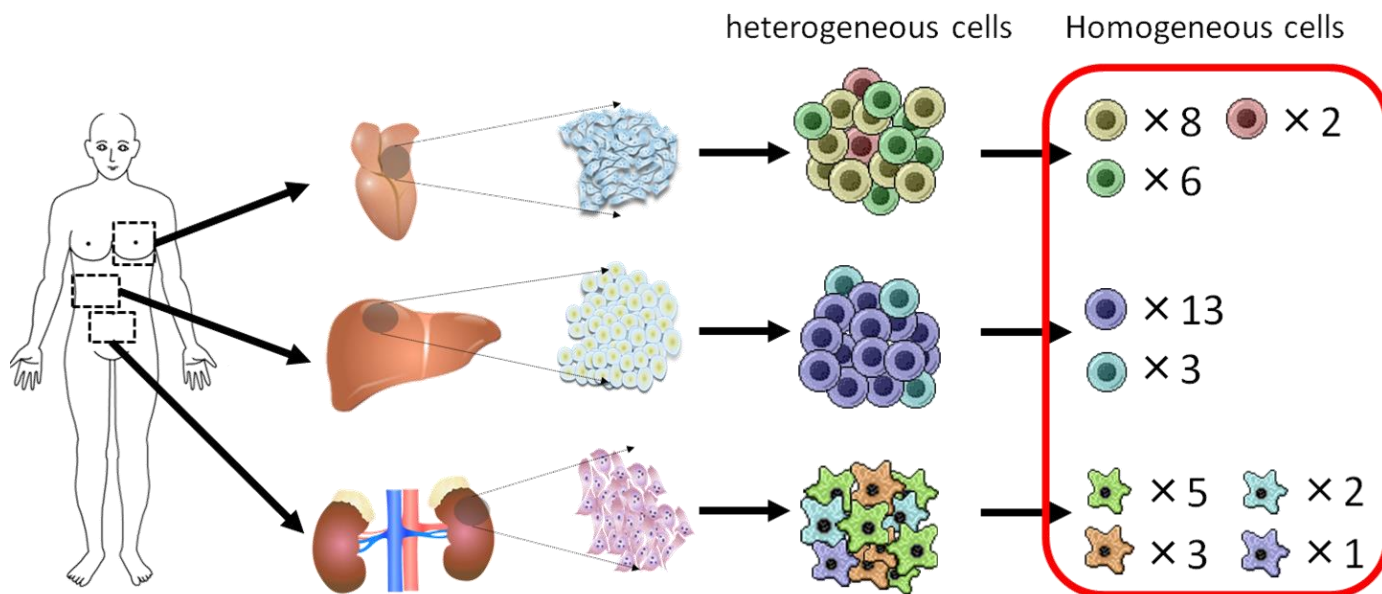
バイオプロセスへの応用

Cyber空間を現実世界へ繋げる



不均一な細胞群データからの情報抽出

細胞不均一性 (Heterogeneity)



Togo picture gallery by DBCLS is licensed under a Creative Commons Attribution 2.1 Japan license (c)

細胞の不均一な制御機構 = 器官形成・多能性に関与

不均一な細胞集団から得られたデータを元に均一な細胞集団の情報のみを抽出する必要がある。⇒**新規クラスタリング技術の開発**

御清聴ありがとうございました

ライフ・テクノロジー
**～生体ビッグデータの解析を通して、健康で
安心・安全な生活を実現する～**

2019年12月17日
「エジソンの会」第34回会合

於：公益財団法人国際高等研究所