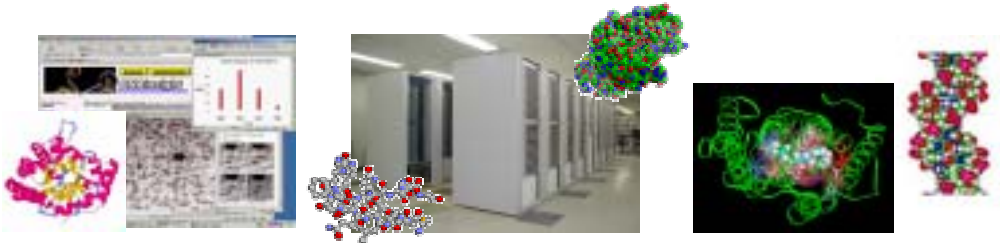



大規模PCクラスタによる バイオインフォマティクス研究



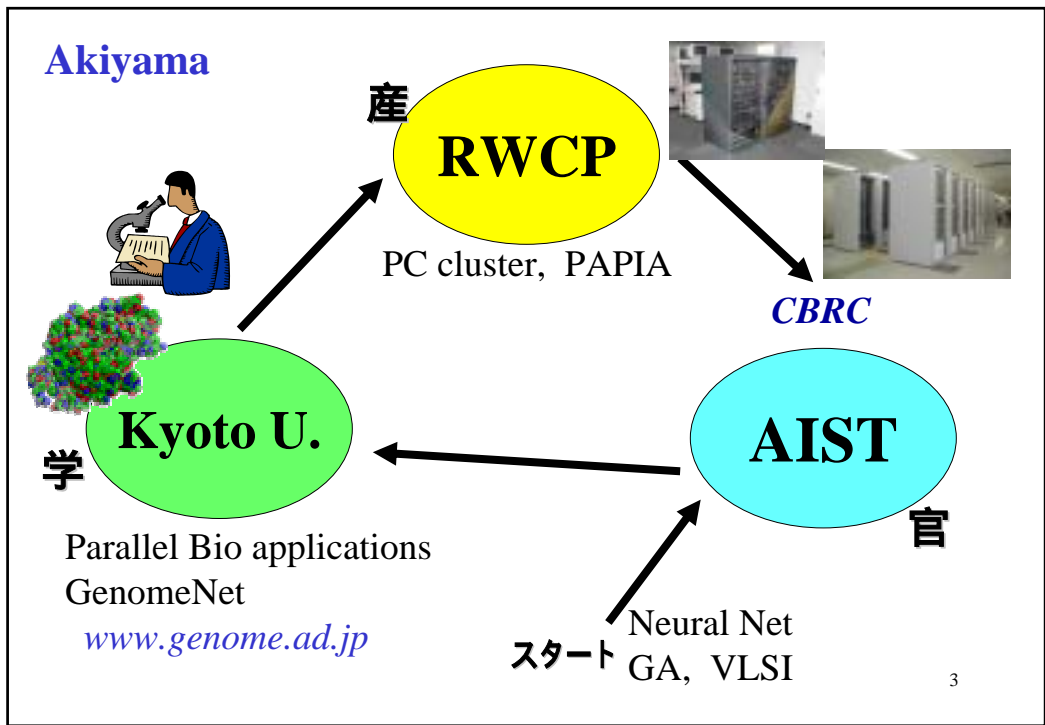
独立行政法人 産業技術総合研究所 
生命情報科学研究センター長 秋山 泰

<http://www.cbrc.jp/>

1

講師略歴

- 中学3年～高校卒業頃まで
 - 大型コンピュータに夢中、計算センター通い
- 大学学部～大学院博士
 - 大規模問題の近似解法(確率的アルゴリズム)
 - 大規模問題を解く並列計算機
- 90-92 電総研 研究官 (バイオへの応用を始める)
- 92-96 京都大学化学研究所 助教授(ゲノムネット)
- 96-00 通産省系 R W C P 研究室長(PCクラスタ応用)
- 00-01 電総研 主研(並列バイオインフォマティクス)
- 01- 生命情報科学研究センター センター長 2



1. バイオインフォマティクス

4



産総研 生命情報科学研究センター

Computational Biology Research Center (CBRC)



2001年4月設立(お台場)

我が国初の産学官連携型
大規模・専門
バイオインフォ研究施設

5



生命情報科学研究センター 組織

センター長 秋山 泰
副センター長 浅井 潔
顧問 深津憲一

基礎系

- アルゴリズムチーム (リーダー 後藤 修 理博)
 - マルチプルアラインメントなど基盤アルゴリズムの効率化等の研究
- 数理モデル・知識表現チーム (リーダー 浅井 潔 工博)
 - 確率モデルによる配列解析、生命情報の格納・検索・演繹方式の研究

応用系

- ゲノム情報科学チーム (リーダー 諏訪牧子 理博)
 - 全ゲノムレベルでの配列解析・機能予測のための手法の開発と応用
- 分子情報科学チーム (リーダー 秋山 泰 工博)
 - タンパク質やペプチドの構造解析・予測のための手法の開発と応用
- 細胞情報科学チーム (リーダー 高橋勝利 理博)
 - 代謝シミュレーションや遺伝子制御ネットの推定等の手法の開発と応用

<http://www.cbrc.jp/>

6

国内有数の人材集結

研究員

人数はH14.12.1現在

産総研正職員	14名
特別研究員等	17名
共同研究員	8名
技術研修員	18名
NEDOフェロー	1名

研究系小計 58名

事務系スタッフ

産総研正職員	2名
非常勤職員	18名
SE契約者	1名



共同研究・技術研修

東京大学・ヒトゲノム解析センター / 奈良先端科学技術大学院大学 / 早稲田大学 / インテック・ウェブ・アンド・ゲノム・インフォマティクス(株) / コンパックコンピュータ(株) / (株)情報数理研究所 / 日鉄日立システムエンジニアリング(株) / NKK(株)(日本鋼管) / 日本電気(株) / (株)富士総合研究所 / 松下電器産業(株) /

7



バイオインフォマティクスとは

● 生命情報科学(バイオインフォマティクス)

幅広い生命現象を、情報論的な立場から扱う

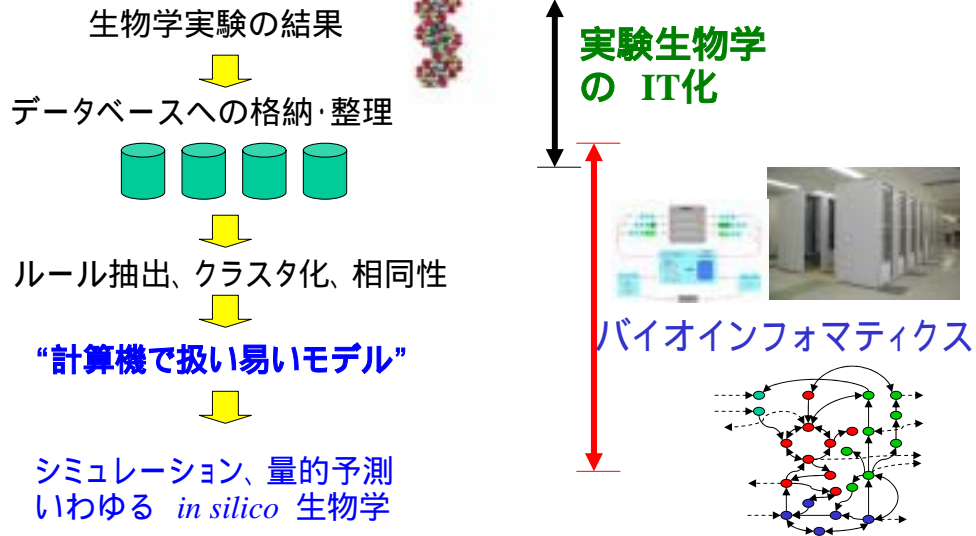
- ゲノム配列の構造、調節の機構
- 産物であるタンパク質分子等の立体構造・機能
- 遺伝子・タンパク質の細胞内での相互作用

● 単にデータベースを作るのは「IT化」に過ぎない

- そのデータベースから計算機でルールをマイニング
- 情報論的に生物を操作する「計算生物学」の勃興

真の「バイオ・IT融合技術」を目指して 8

バイオインフォマティクス (生命情報科学)



バイオインフォマティクスと産業

- ライフサイエンスと情報科学との融合による技術
- 21世紀のバイオ産業を支える「マザー産業」
 - 例)自動車産業発展における、工作機械産業の役割

市場規模(2002年推定):

バイオテク関連: 米国7.4兆円、欧州3.5兆円、日本1.3兆円
 内、バイオインフォ: 米国680億円、欧州260億円、日本100億円
 (出典: 特許庁「平成12年度特許出願技術動向調査分析報告書」)

インフォ用計算機・データベースは既に2003年世界で90億ドル。
 (出典: 米国IBM社による推定。日本経済新聞 2001.7.25)

国家BT戦略 2010年国内バイオテク市場 10~25兆円。
 2兆円が「**バイオツール**」、2兆円が「**バイオインフォ**」

バイオインフォマティクスへの期待

クリントン-ブレア：ヒトゲノム配列読取終了宣言

『人類が作った最も重要な地図』（2000年6月26日）

遺伝子の**意味解釈**の国際競争が激化。



バイオインフォマティクス（生命情報科学）

- **期待1（高速性）**

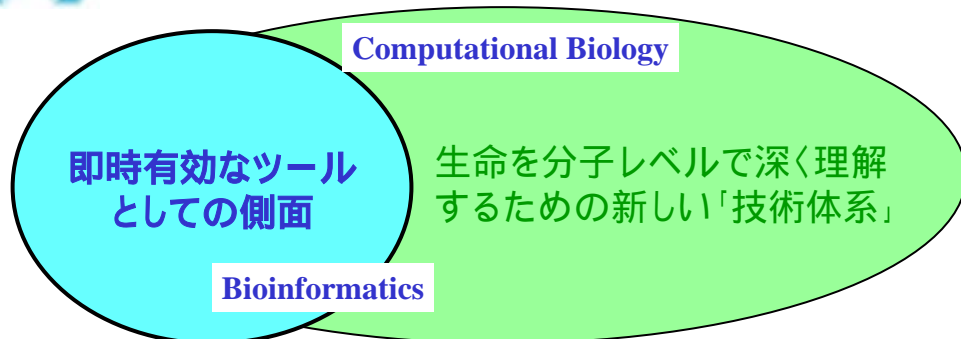
膨大情報からいち早く産業的価値を発見、特許化。

- **期待2（経済性、安全性、倫理性）**

コンピュータで候補絞り込み、試行錯誤の実験減らす。



生命情報科学の2つの使命



遺伝子発見
機能予測



ゲノム情報の早急な産業利用
国際貢献、知的財産権の確保

1～5年以内に明確な貢献

細胞モデリング
相互作用予測
分子デザイン



試行錯誤的な実験を減らす
コスト削減、時間短縮、安全・倫理

21世紀のバイオ産業の基盤技術

外的動向：海外のバイオインフォマティクス研究所

- 米国 NIH傘下

- NCBI (National Center for Biotechnology Information)

- 1988年設立 現在約300名

- 実験との独立、計算機に特化。
数年内の500名体制を宣言して、
ポストク倍増。David Lipman所長



創設者
Claude Pepper上院議員



NIH-Bldg. 38A



NIH-Bldg. 45

- 欧州連合 EBI (European Bioinformatics Institute)

- 1992年設立 現在約200名

- 英国ケンブリッジ郊外。Janet Thornton所長



EBI

さらにポーランド、シンガポール、韓国等
が急激に実力を付けている。各国内に複数の中核拠点。

日本のバイオインフォマティクス研究拠点

- ▶ 日本DNAデータバンク (DDBJ) S61 ~
- ▶ 東大医科研・ヒトゲノム解析センター (HGC) H3 ~
- ▶ 理研・ゲノム科学総合研究センター (GSC) H11 ~
- ▶ 京大化研・バイオインフォマティクスセンター H13 ~
- ▶ 産総研・生命情報科学研究センター (CBRC) H13 ~
- ▶ 産総研・生物情報解析研究センター (BIRC) H13 ~

他

バイオインフォマティクスの大規模・専門研究所が必要。
インフォマティクスの基礎から応用まで、人材を集結。

私の視点1

事実： バイオインフォマティクスは 新たな実験手法にドライブされてきた

DNAシーケンサ
 全ゲノムショットガン法
 全ゲノムシーケンシング
 X線、NMR等
 マイクロアレイ、チップ
 質量分析
 SNPs解析
 細胞イメージング
 ナノテク、 μ TAS

相同性解析、進化系統樹
 アセンブリ
 比較ゲノミクス
 立体構造解析、予測
 発現解析、ネットワーク推定
 タンパク質同定、相互作用解析
 薬剤感受性等との連鎖解析
 細胞内位置依存の -omics

私の視点2

「整理」のバイオインフォ

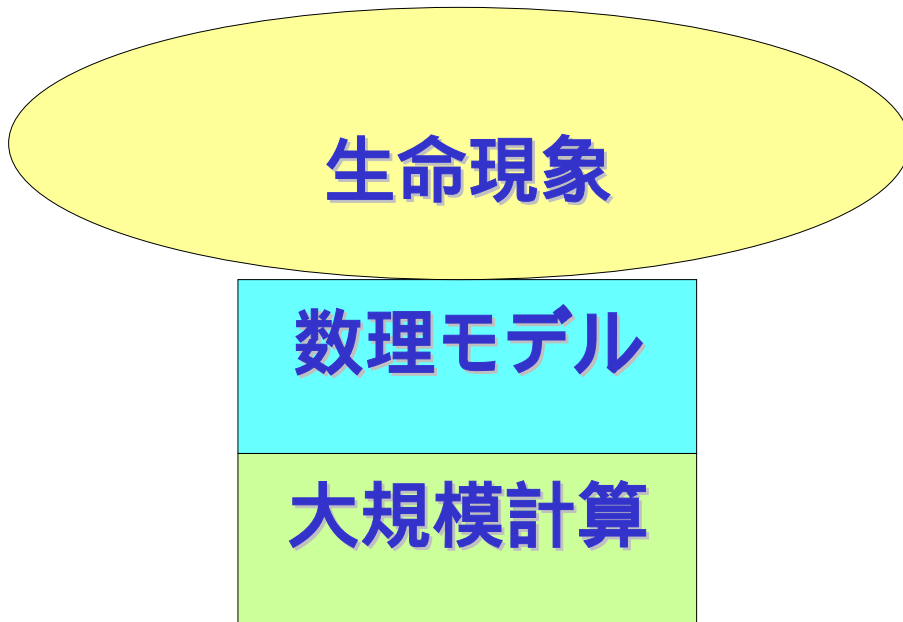


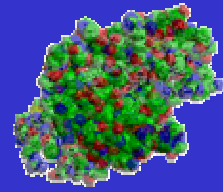
「変換」のバイオインフォ



「模擬」のバイオインフォ

私の視点3
生命現象
数理モデル
大規模計算

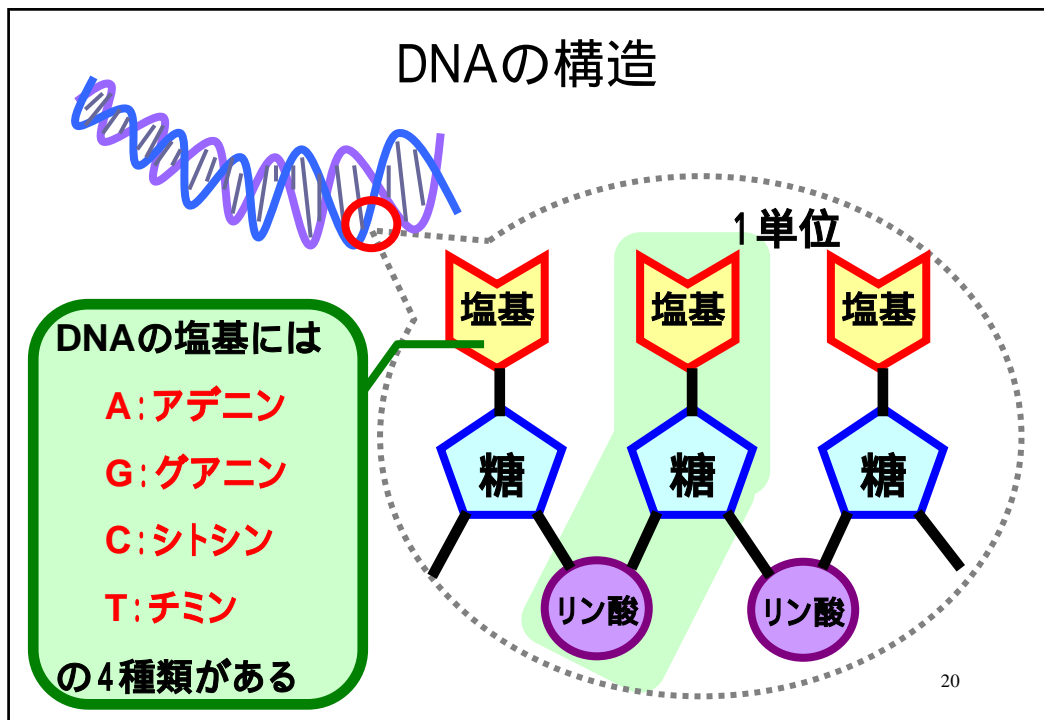




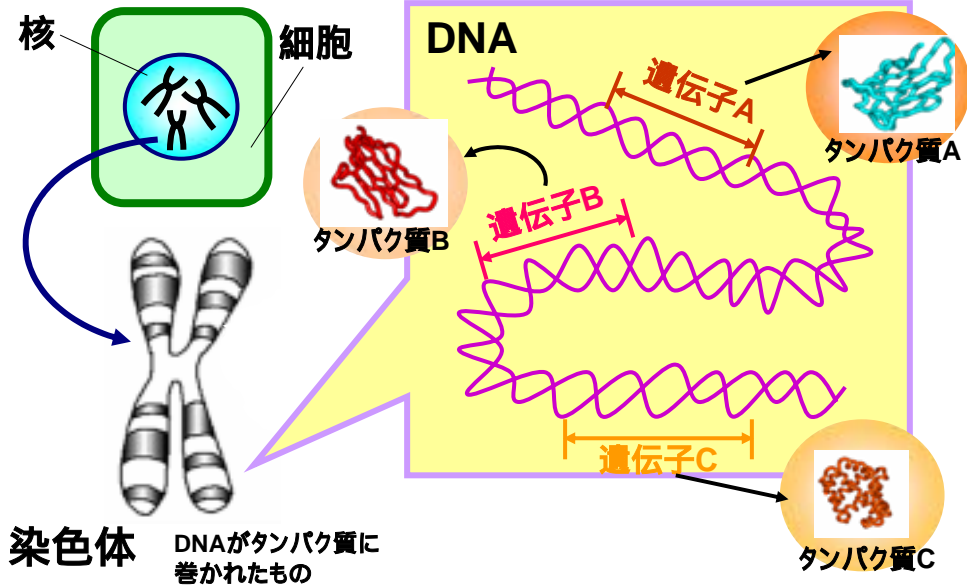
2. ゲノム解析の衝撃



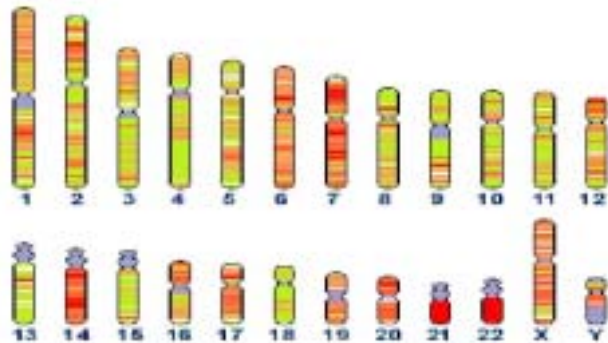
19



遺伝子にはタンパク質を作るための情報が記されている



ヒトのゲノム



- ・ヒトの細胞には、23対の染色体
(22対の常染色体、1対の性染色体)
- ・これらの上に書かれている、
遺伝情報の総体 = ゲノム (genome)

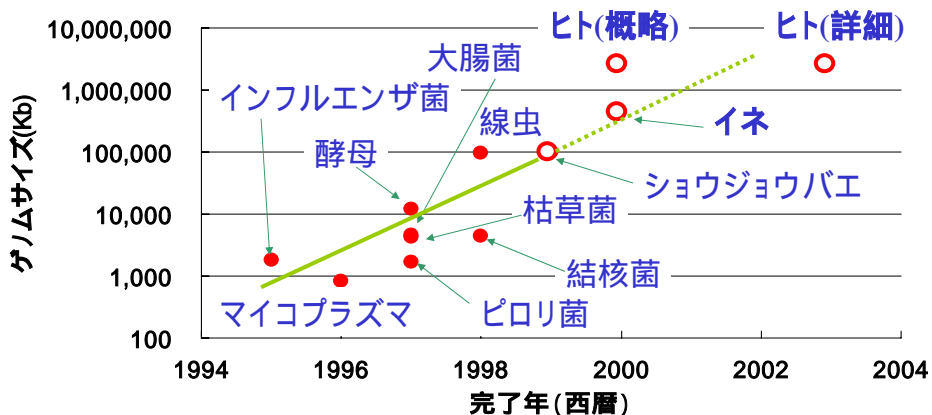
もしも、ヒトのDNAを まっすぐに延ばせたらどの位の長さ？



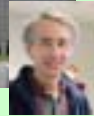
1番染色体	3億文字	10.3cm	13番染色体	9800万	3.4cm
2番染色体	2億5500万	8.7cm	14番染色体	9300万	3.2cm
3番染色体	2億1400万	7.3cm	15番染色体	8900万	3.1cm
4番染色体	2億0300万	7.0cm	16番染色体	9800万	3.4cm
5番染色体	1億9400万	6.7cm	17番染色体	9200万	3.2cm
6番染色体	1億8300万	6.3cm	18番染色体	8500万	2.9cm
7番染色体	1億7100万	5.9cm	19番染色体	6700万	2.3cm
8番染色体	1億5500万	5.3cm	20番染色体	7200万	2.5cm
9番染色体	1億4500万	5.0cm	21番染色体	3383万	1.2cm
10番染色体	1億4400万	5.0cm	22番染色体	3460万	1.2cm
11番染色体	1億4400万	5.0cm	X染色体	1億6400万	5.6cm
12番染色体	1億4300万	4.9cm	Y染色体	3500万	1.2cm

1番から22番は2本ずつ、男性はXとY、女性はXを2本、合計するとおおよそ2m。

ゲノム解析の進展



世界のゲノム解析施設とコンピュータ



イギリス ケンブリッジ
Sangerセンター

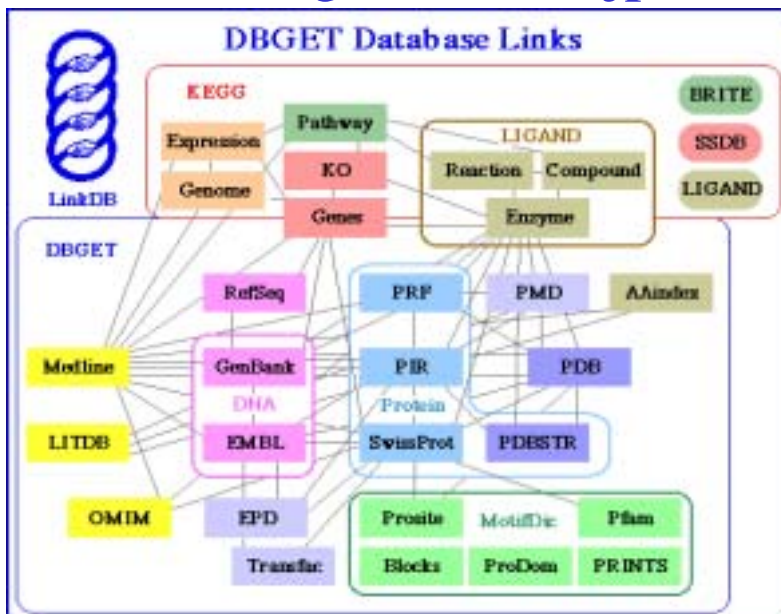
アメリカ セントルイス
ワシントン大学



アメリカ メリーランド州
Celera Genomics社






www.genome.ad.jp



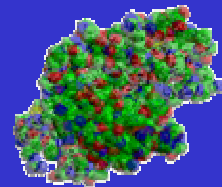
参考： データ数のギャップ

- DNA seq. are rather inexpensive information
- 3-D structures are very expensive information

	type	DB	version	#entries	
	DNA	nr-nt	02-06-25	17563K	← 1/20
	Amino	nr-aa	02-06-25	941K	← 1/50
	3-D	pdb	02-06-24	18K	



27

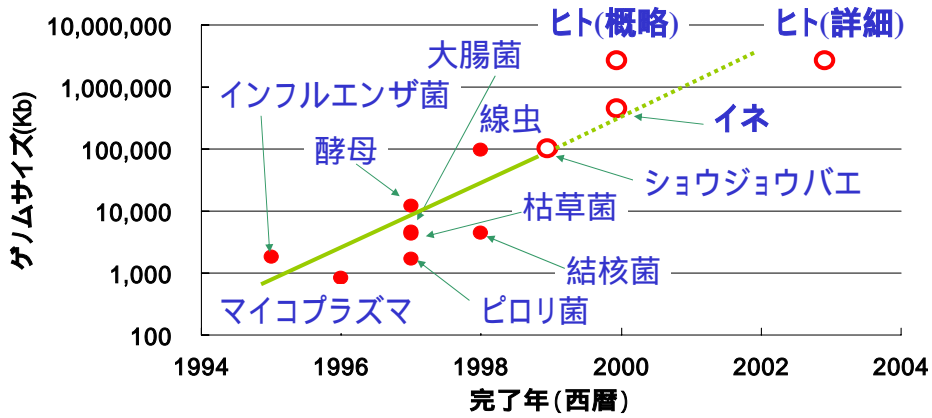


3 . バイオインフォにおける 並列処理技術の浸透



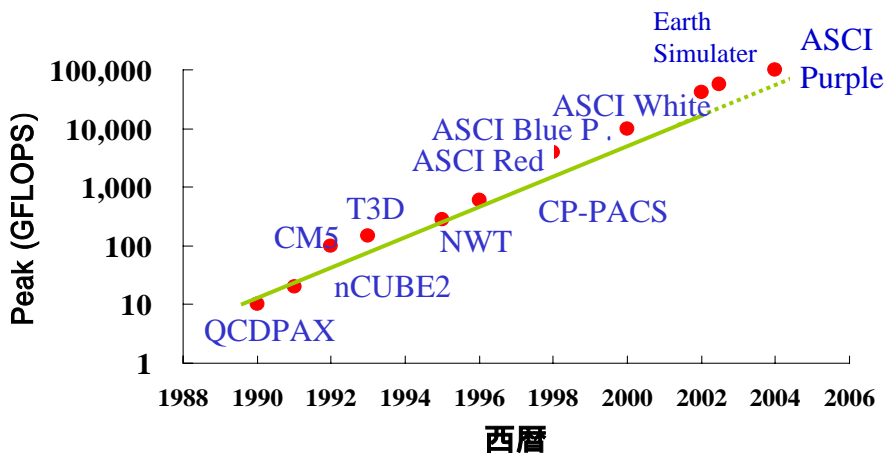
8

ゲノム解析の進展



29

並列計算機のピーク性能の進歩



バイオインフォマティクス における多様な並列性

- データベース分割による並列性
- 木探索の並列性
- 確率的試行の並列性
- データ内部の細かい並列性

文献

秋山泰:大規模並列処理によるタンパク質情報解析,
人工知能学会誌, Vol.15, No.1, pp.27-34, (2000).

31

1) データベース分割による並列性

- データベースを分割して複数プロセッサで処理
 - 配列相同性検索 [Barton90, 五條堀91ほか]
 - 配列モチーフ検索
 - 構造類似性検索
 - PAPIAシステム(秋山ら)
 - PDB構造との総当たり構造比較
 - 配列や構造のクラスタリング
 - PDB-REPRDB(野口ら)

負荷分散のよいデータベース分割法がポイント

32

2) 探索空間の分割による並列性

- 最適解探索において、膨大な探索空間を分割
 - 並列反復改善マルチプルアライメント[広沢95]
 - 最尤法に基づく進化系統樹作成[松田94]
 - 決定木によるモチーフ発見[Shoudai 95]
 - データマイニングにおける知識発見[森下99]
 - 確率文法による配列特徴発見[馬見塚94]
 - RNAの二次構造予測[竹田92]
 - ペプチド可能配座解析[安藤99]

枝刈りにともない、部分空間の効率的な再割当が必須

33

3) 確率的試行の分割による並列性

- 乱数系列や初期値を変更して、独立に実験。
- ほとんど通信は要らない。2)の特殊例。
 - 並列GAによるマルチプルアライメント[戸谷95]
 - フォールディング予測[Skolnick]
 - ほか

**実現は極めて容易。低速ネットでもよい。
ただし、計算の途中結果を活かしたり、結果に応じて
試行の方向をステアリングする場合は、通信が必要。**

34

4) データ構造の細分割による並列性

- 長いDNA配列、大きなタンパク質分子等を分割。
- 並列化アルゴリズムの構築が必要
 - 相同性解析 [Brutlag93, Jones92] CM2, Maspar等
 - タンパク質の露出表面積 [Martino94]
 - 分子動力学法計算
 - MolTreC – カットオフをしない並列分子動力学法
 - 分子軌道法計算

1データだけの計算でも高速化可能になる。
ただし労も多く、より安易な並列性が使えるときは無用か。

35

PAPIA cluster (1998)



CPU数: 64
Node数: 64
CPU: Pentium Pro
200MHz
Memory: 16GB
Disk: 256GB
OS: NetBSD 1.2.1
並列OS: SCore 2.0

PAPIA サービスに特化 (インターネット上で24時間無料計算サービス)

PAPIA (Parallel Protein Information Analysis) system
<http://www.cbrc.jp/papia/> 並列タンパク質情報解析システム

36

PAPIA mini cluster (1999)



CPU数: 16
Node数: 8
CPU: Pentium II
450MHz, Dual
Memory: 4GB
Disk: 50GB
OS: RH Linux 2.2.3
並列OS: [SCore 3.0](#)

移動可能なためデモンストレーション用などに利用

LUCIEプロジェクト(東工大)用のテスト環境
Open BioGRIDプロジェクト用のテスト環境、等

37

Compaq cluster (2000)



CPU数: 64
Node数: 32
CPU: Pentium III
1.4GHz, Dual
Memory: 32GB
Disk: 1152GB
OS: RedHat Linux 6.2
並列OS: [SCore 3.2](#)

分子動力学法計算 MolTreC などの開発環境
分子軌道法計算 ABINIT-MP などの開発環境
タンパク質質量分析計算、他
(他チームによる利用では、遺伝子発見、単粒子解析、他)

38

Magi system (2001)

Massively Parallel Computer
for Genome Informatics



CPU数: 1040
Node数: 520
CPU: Pentium III
933MHz, Dual
Memory: 520GB
Disk: 19152GB
OS: RedHat Linux 7.1
並列OS: SCore 4.1

分子動力学法計算 MolTreC, AMBER
分子軌道法計算 ABINIT-MP, Gaussian
非冗長タンパク質DB作成 PDB-REPRDB
タンパク質相同性総当たり計算、他
(他チームによる利用では、遺伝子発見、単粒子解析、他)

2001年秋時点では、
32bit 級PCクラスタで
Linpack実効性能世界一
653.8 GFLOPS
(ピーク性能の67%)

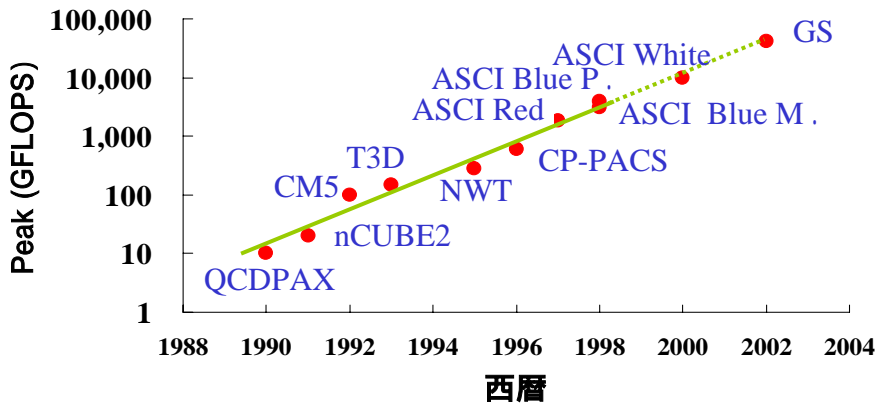
スパコンTOP500 <http://www.top500.org>

順位	名称	実測性能 (GFlops)	理論最大 (GFlops)	PU数	設置機関
1	ASCI White	7226	12288	8192	米国立ローレンスリバモア研
2	Compaq ES45	4059	6048	3024	米ピッツバーグスパコンセンター
3	IBM SP3	3052	4992	3328	米エネルギー研究科学計算センター
...					
39	CBRC Magi Cluster	654	970	1040	産業技術総合研究所
40	SCoreIIIe Cluster	618	955	1024	新情報処理開発機構
41	IBM P Cluster	594	1024	1024	米国立スパコン応用センター (NCSA)

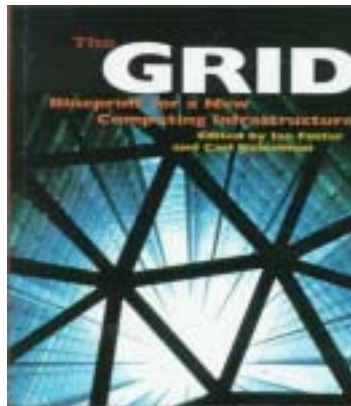
2002年11月は86位

古い！！ 2001年11月版ランキング

並列計算機のピーク性能の進歩



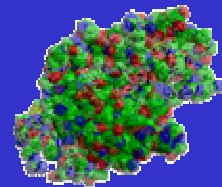
GRID



- Networkは到達性、
Gridは集団機能を提供
- (少なくとも) 3つの異なるレベル、形態
 - メタコンピューティング Folding@home
 - PCクラスタで超スパコン (合わせて性能発揮)
 - 仮想スパコンセンター (物理的でなく論理的対応)

1000 : 1

1 year : 9 hours



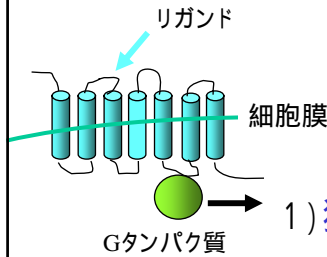
**4. さて、1000倍速いと
何が出来る**



研究事例1:創薬標的遺伝子の自動発見

Gタンパク質共役型受容体(GPCR)

既存薬の約半数は、GPCR分子が標的。

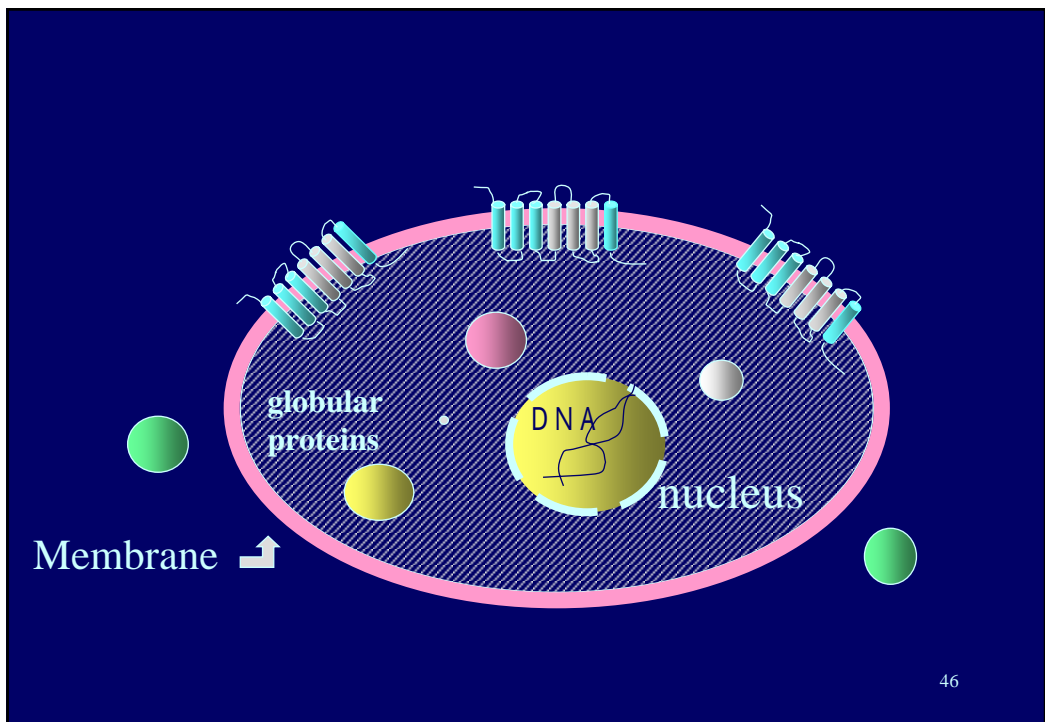


- 1) 独自理論(多重出力HMM法)の適用
10~15エキソンに細かく分断されていても発見。浅井ら
- 2) 膜タンパク質情報解析の技術蓄積
7回膜貫通型タンパク質を発見するノウハウ。諏訪ら
- 3) 大規模並列計算により精密計算が可能
1台で32週、250台で20時間、1000台で5時間。秋山ら



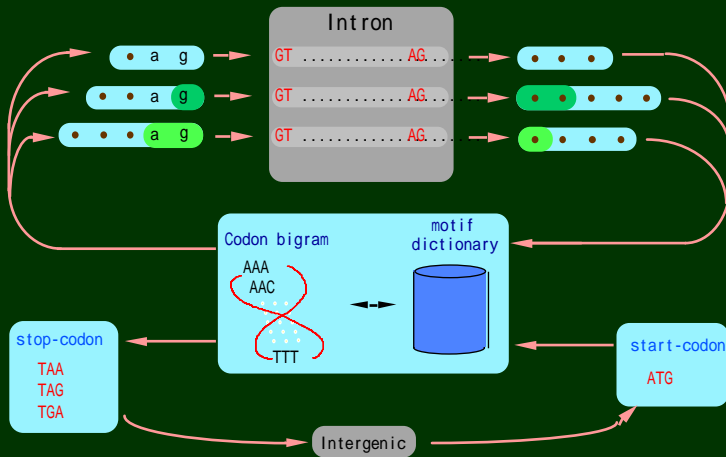
国内大手企業と、薬剤標的同定の共同実験へ

45



遺伝子領域予測システム GeneDecoder

<http://www.genedecoder.org/>



47

遺伝子発見 (Gene Finding)

- 目的
 - DNA配列から遺伝子の部分を**コンピュータ**で発見
- 手がかり
 - 配列上のシグナル
 - コード領域らしさ (Coding Potentials)
 - 相同性 (ホモロジー)

48

“6種の読み枠 で 翻訳してみる”



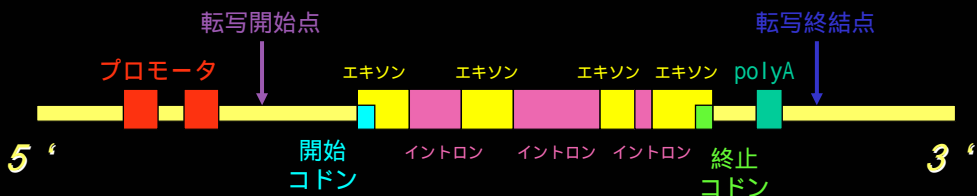
Translate until **STOP** codon with 6 possible frame

Extremely easy application (but useful)

$O(n)$

49

遺伝子の構造 (真核生物)



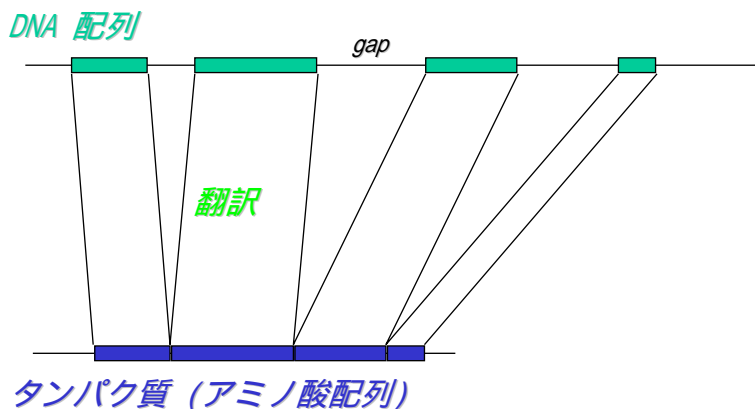
遺伝子は何処？

- 遺伝子を探そう !!

```
tttgaataaagaggcgtgccttccaggcaggctctataagtgaccgccgaggcgagcgt
gcgcgcgttgcaggtcactgtagcggacttctttgggtttctttctctttggggcacct
ctggactcactccccagcatgaaggcgctgagcccgggtgcgcggctgctacgaggcggtg
tctgctgctgtcggaacgcagtctggccatcgccggggccgagggaggccggcagct
gaggagccgctgagcttctggacgacatgaaccactgctactcccgcctgcgggaactg
gtaccggagctccgagaggcactcagcttagccagggtggaaatcctacagcgcgtcatc
gactacattctcgacctgcaggtagctctggccgagccagccccggacccccgtatggc
ccccaccttcccatccaggtaagcctcgaagtctgggacagggctgaacaccaggcaagg
atgctgcgggacctcggagctcccgatgcctcgcgtaactcttccctcttttctctta
atcagacagccgagctcgtccggaactgtcatctccaacgacaaaaggagcttttggc
actgactcggccgtgtcctgacacctccaggtagtatctctctcttggagagggaggt
ttaaacggcaagtctggagtggcagacgttttgaaaaatgcccactcactcggtttag
```

51

イントロンがあるから難しい



52

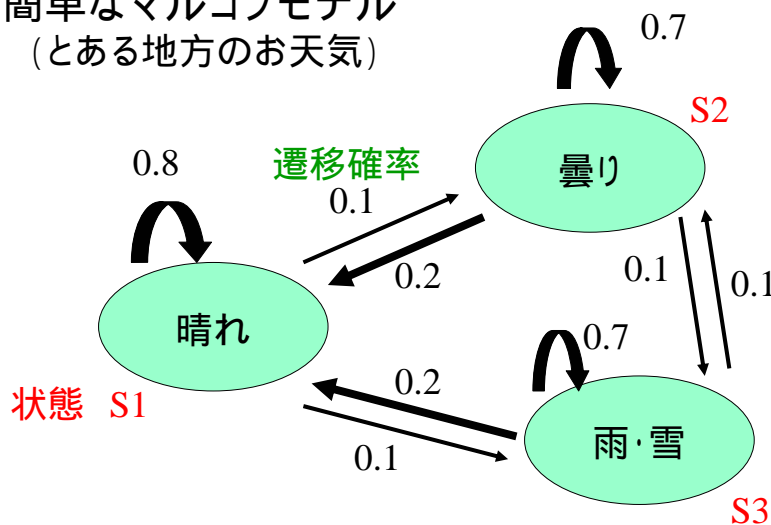
これが正解でした！

- Human, Id3 gene for HLH type transcription factor

```
t t t g a a t a a g a g g c g t g c c t t c c a g g c a g g c t c T A T A A g t g a c c g c c g c g g c g a g c g t
g c g c g c g t t g c a g g t c a c t g t a g c g g a c t t c t t t g g t t t t c t t t c t c t t t g g g g c a c c t
c t g g a c t c a c t c c c c a g c A T G a a g g c g c t g a g c c c g g t g c g c g g c t g c t a c g a g g c g g t g
t g c t g c c t g t c g g a a c g c a g t c t g g c a t c g c c c g g g c c g a g g g a a g g g c c c g g c a g c t
g a g g a c c g c t g a g c t t g c t g g a c g a c a t g a a c c a c t g c t a c t c c c g c c t g c g g g a a c t g
g t a c c c g g a g t c c c g a g a g g c a c t c a g c t t a g c c a g g t g g a a t c c t a c a g c g c g t c a t c
g a c t a c a t t c t c g a c c t g c a g g t a g t c c t g g c c g a g c c a g c c c t g g a c c c c t g a t g g c
c c c c a c c t t c c c a t c c a g g t a a g c c t c g a a g t c g g g a c a g g g c t g a a c a c c c a g g c a a g g
a t g c t g c g g g a c c c t c g g a g c t c c c g a t t g c c t c g c g t a a c t c t t c c c t c t t t t c c t c t a
a t c a g a c a g c c g a g c t c g c t c c g g a a c t t g t c a t c t c c a a c g a c a a a a g g a g c t t t t g c c
a c T G A c t c g g c c g t g t c c t g a c a c c t c c a g g t g a g t a t c t c c t c t c t t g g a g a g g g a g g t
t t a a a c g g c a a g t c c t g g a g t t g g c a g a c g t t t t g a a a a t t g c c a c t c a c t c g g t t t a g
```

53

簡単なマルコフモデル (とある地方のお天気)



一次のマルコフモデル (1つ前の過去だけに依存) 54

ちなみに、定常分布を求めてみると。。。

$$\begin{pmatrix} P1' \\ P2' \\ P3' \end{pmatrix} = \begin{pmatrix} 0.8 & 0.2 & 0.2 \\ 0.1 & 0.7 & 0.1 \\ 0.1 & 0.1 & 0.7 \end{pmatrix} \begin{pmatrix} P1 \\ P2 \\ P3 \end{pmatrix}$$

遷移確率行列

$$P1+P2+P3 = 1.0$$

$P1'=P1$, $P2'=P2$, $P3'=P3$ として連立方程式を解くと

$$P1 = 0.5, \quad P2 = 0.25, \quad P3 = 0.25$$

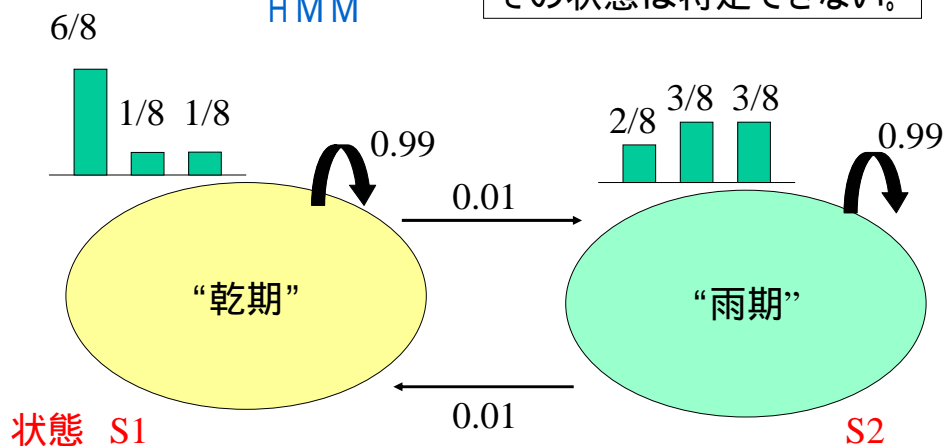
微妙な遷移確率の差だったが、
晴れの日が支配的になっている。

55

隠れマルコフモデル

HMM

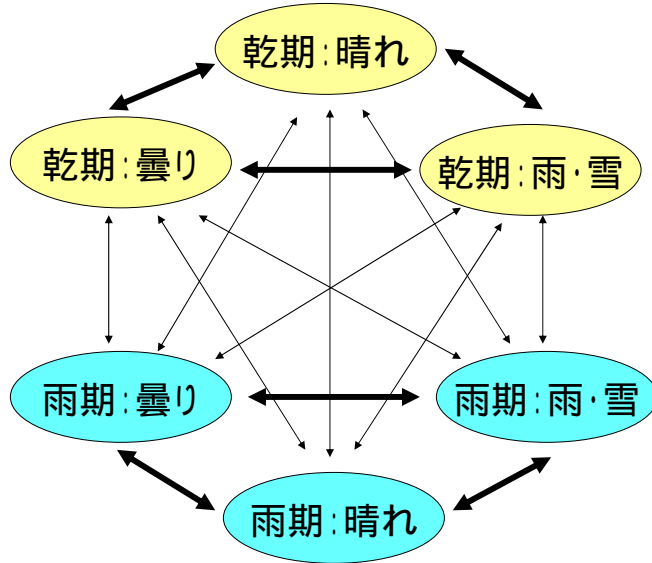
出力信号は“晴れ”でも、
その状態は特定できない。



各状態が3種類の出力{晴、曇、雨}を確率的に出力

56

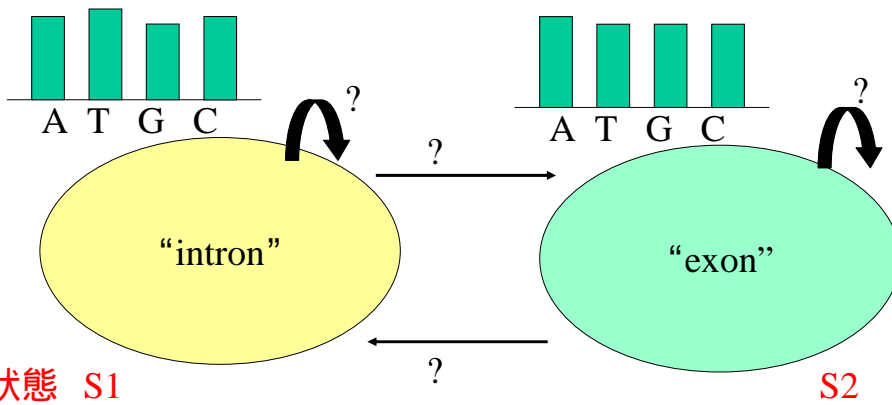
ちなみに、(旧状態数2) × (出力文字数3) に展開すれば、



出力信号が単種のモデルに戻せるが、パラメータ多くなる。⁵⁷

隠れマルコフモデル
(遺伝子構造解析へ?)

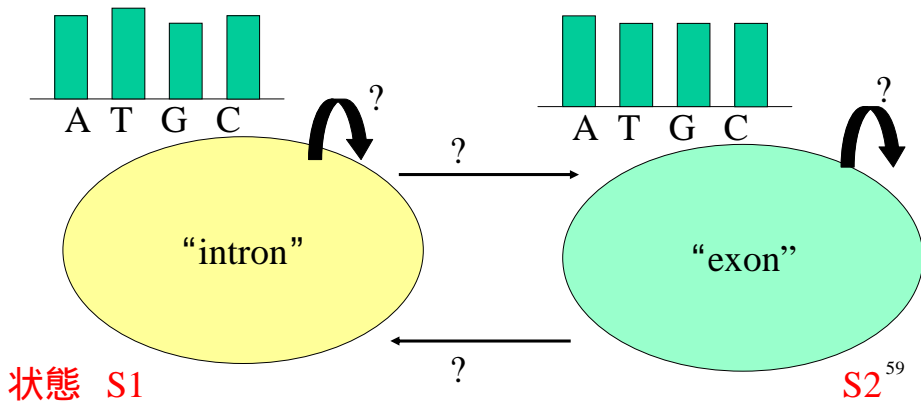
出力信号{A,T,G,C}から、
その隠れ状態は特定できない。



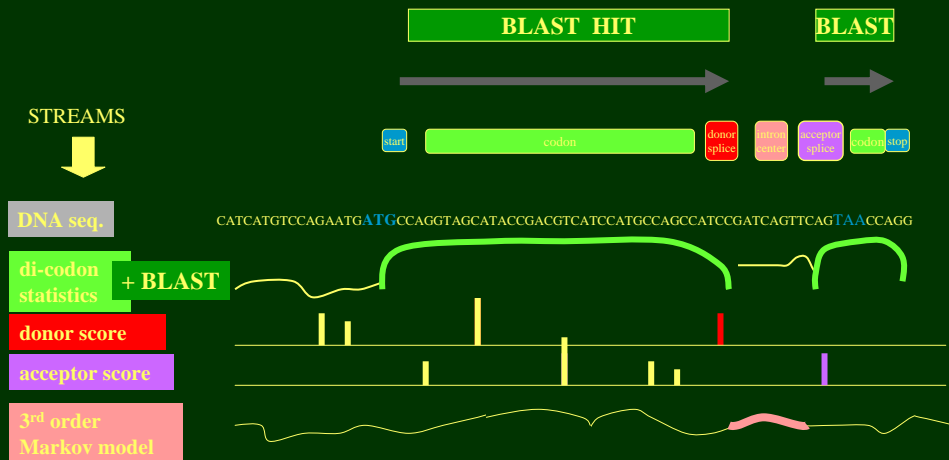
出力信号に、{A,T,G,C}を使うだけでは弱いが、
3文字(コドン)や、6文字を利用すれば偏りが見られる?⁵⁸

隠れマルコフモデル

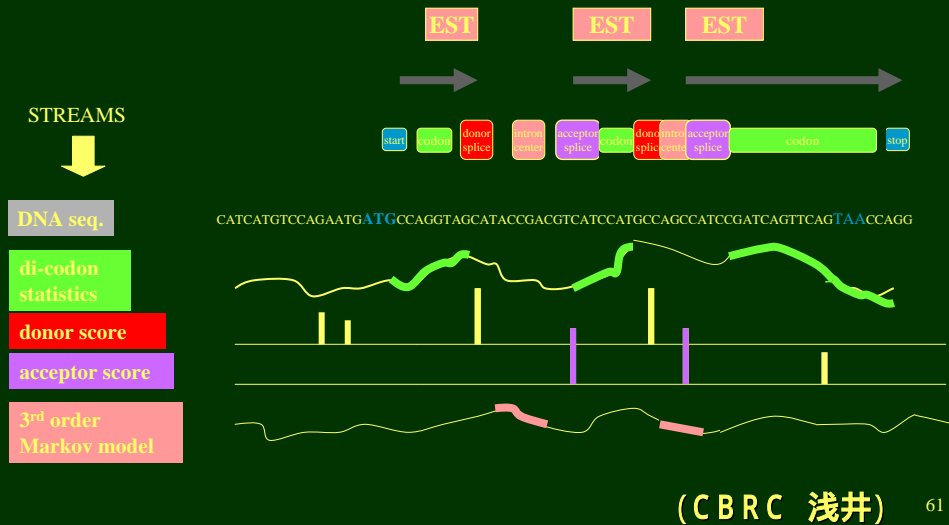
- 1 a. 沢山のデータから、モデル上の確率パラメータを推定
- 1 b. 最適なモデル(形状)を推定
- 2. 与えられた新規「出力信号列」から、状態遷移を逆推定



BLAST相同性検索の統合



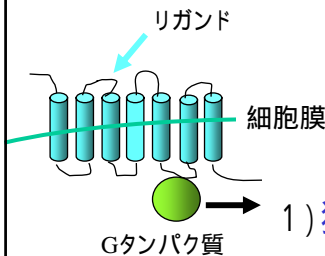
ESTデータの統合



研究事例1: 創薬標的遺伝子の自動発見

Gタンパク質共役型受容体(GPCR)

既存薬の約半数は、GPCR分子が標的。



- 1) 独自理論(多重出力HMM法)の適用
10~15エキソンに細かく分断されていても発見。浅井ら
- 2) 膜タンパク質情報解析の技術蓄積
7回膜貫通型タンパク質を発見するノウハウ。諏訪ら
- 3) 大規模並列計算により精密計算が可能
1台で32週、250台で20時間、1000台で5時間。秋山ら



国内大手企業と、薬剤標的同定の共同実験へ

2000 “Druggable” genes?

3 ~ 5万個のヒト遺伝子の中で、市場性のある創薬標的遺伝子は2000個前後という説。グラクソや武田等の大手はやり尽くしたと宣言したが、まだ技術的な穴がある。高度解析で食い込む余地大。

- 738 GPCR **我々の予想: 850個前後?**
- ~ 450 プロテアーゼ
- > 450 キナーゼ
- ~ 300 イオンチャンネル
- 48 核レセプター
- 22 インテグリン

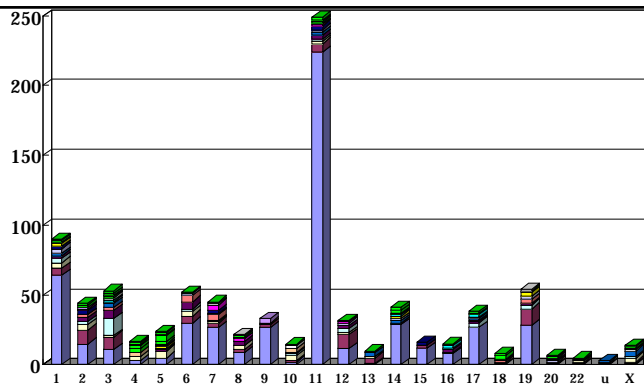
2001年9月

主な創薬標的遺伝子の数
Dr Steve Foord (英国GSK)

ヒトゲノム上の GPCR 分布

11番染色体
に多くの候補

ただし水色表示は
臭覚等レセプター



Automatic discovery of G-protein coupled receptors (GPCR) from human genome draft sequences

The screenshot displays the SEVENS web interface. The main heading is "SEVENS" in a stylized font. Below it, the text reads "Result of Content Search" and "558 sequences found." A table lists search results with columns for ID, Accession, and Name. A detailed view of a sequence is shown in the foreground, including the "ORF sequence" and "Amino acid sequence". To the right, a 3D ribbon diagram of a GPCR protein structure is visible, colored in green and blue.

65



大規模計算により、可能に

コンピュータ
の台数

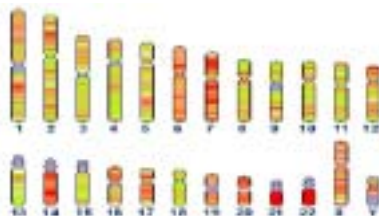
ヒトゲノム配列を扱った場合

1台
32台
1000台

32週
7日
5時間



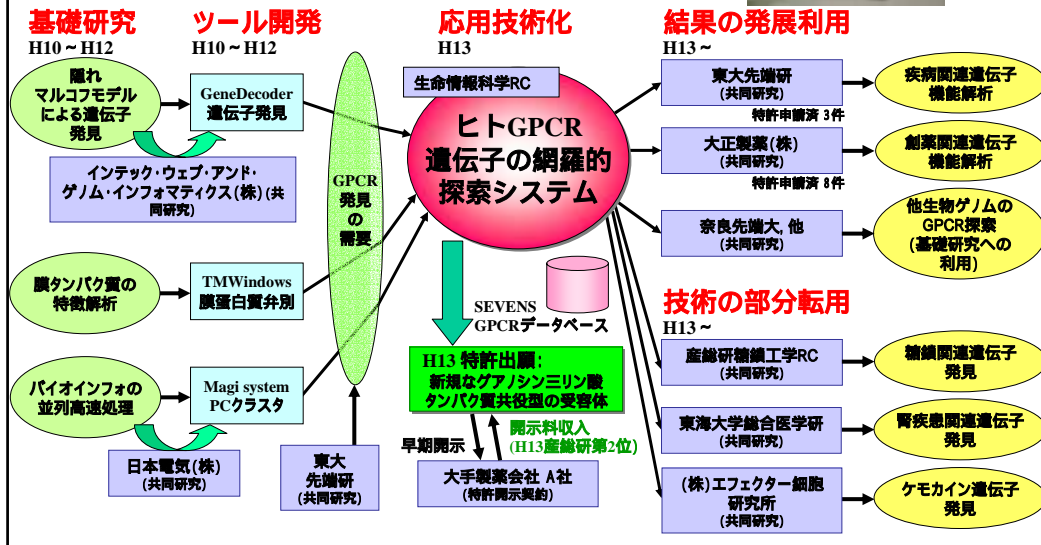
Magi



66

図II-3-2 産総研ライフサイエンス分野の特徴

平成13年度の具体的成果(例)
ヒトGPCR遺伝子の網羅的探索システム



麹菌 *Aspergillus Oryzae*
ゲノム解析コンソーシアム

CBRCは遺伝子発見を担当。
計算機で自動解析、特許申請。

町田グループと共に
コンソーシアム参加

日本経済新聞
2001.9.24

研究事例2： 単粒子解析

SPA: Single Particle Analysis



電圧感受性Naチャンネルの
立体構造を決定

ヘリウム温度凍結溶液を電顕で撮影

単分子画像の情報解析

Nature 2001年2月22日号掲載
AIST Today 創刊号にも紹介

Sato, Ueno, Asai 5

69

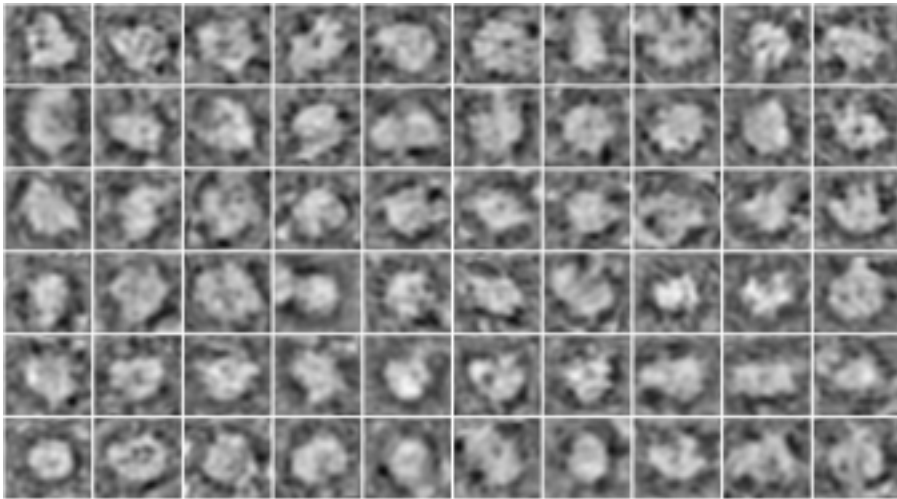
単粒子画像のピックアップ



画像提供：佐藤主税、上野豊(産総研)

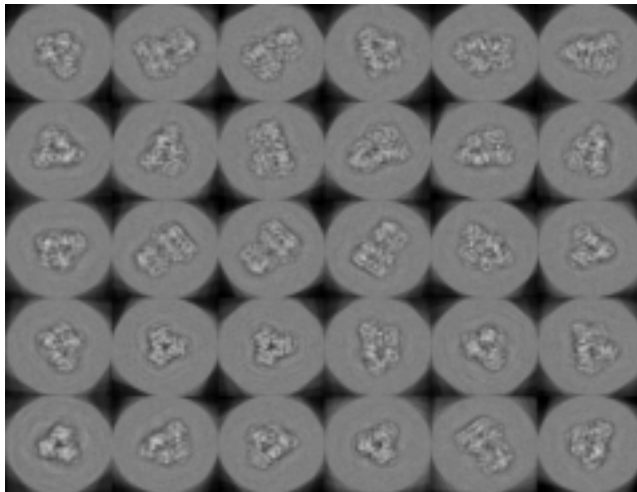
70

example protein SPA images



C. Sato and Y. Ueno₇₁

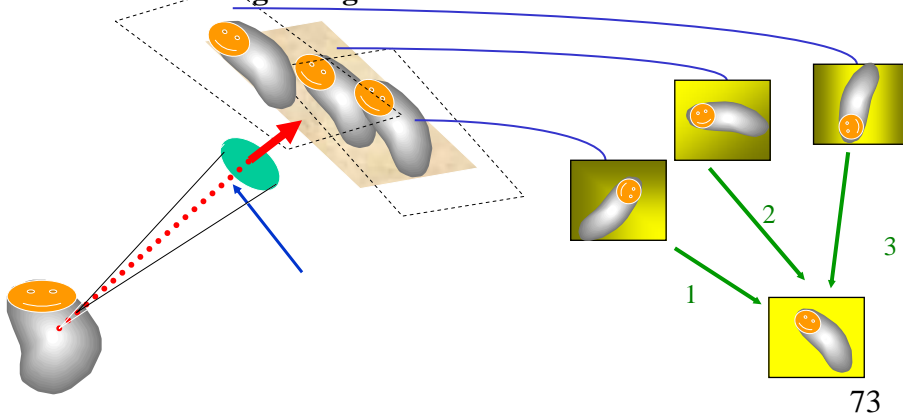
2-D image clustering and averaging



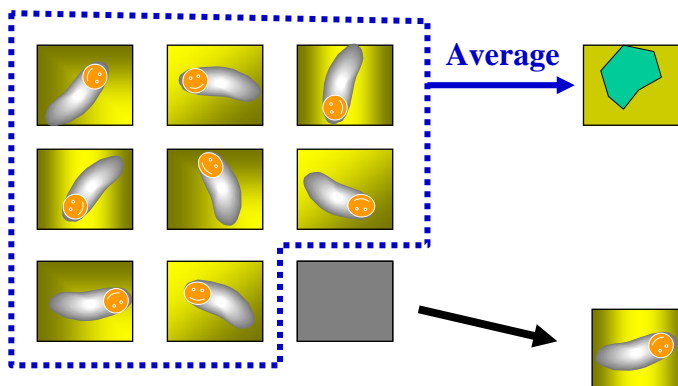
C. Sato and Y. Ueno₇₂

image projection and clustering

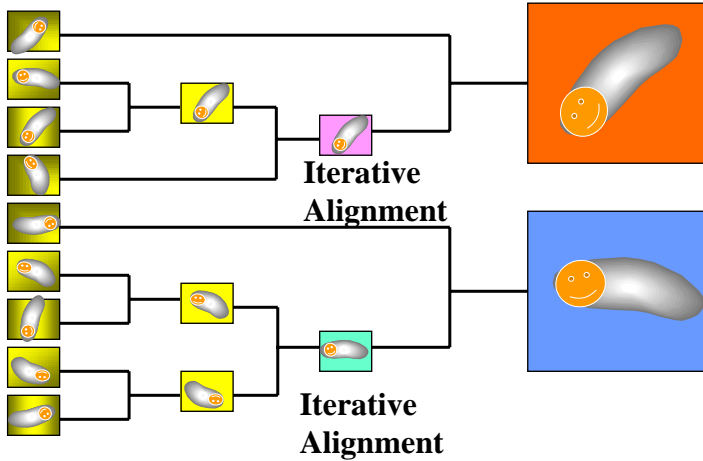
- aim
 - to find images with similar i , i
 - to find i
 - to make an average image



Iterative Alignment without reference



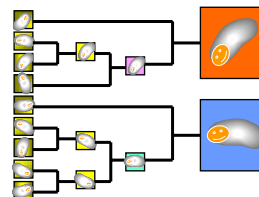
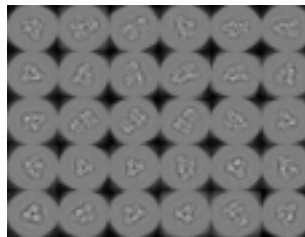
ボトムアップクラスタリング



75

画像のボトムアップクラスタリング

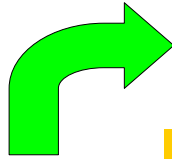
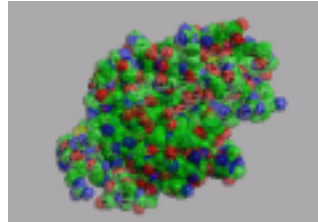
コンピュータ の台数	電子顕微鏡画像の個数	
	1000	20000
1台	1週間	12年
32台	6時間	4月
1024台	10分	2日



画像提供: 佐藤主税、上野豊 (産総研)

研究事例3

立体構造



配列情報

M-K-A-L-I-V-L-G-L-V

タンパク質立体構造予測

並列処理技術

77

1100台のPCで折れ畳み予測

■ Jeff. Skolnick グループ

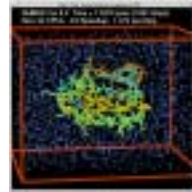
- ▶ Danforth 植物科学センター (アメリカ St. Louis)
<http://www.danforthcenter.org/> **現在はバッファロー大学**
- ▶ Dr. Skolnickのタンパク質構造予測グループ。 **確率試行並列**
- ▶ Dual P-III 731MHz × 490 + Dual P-II 400MHz × 60で約\$2M



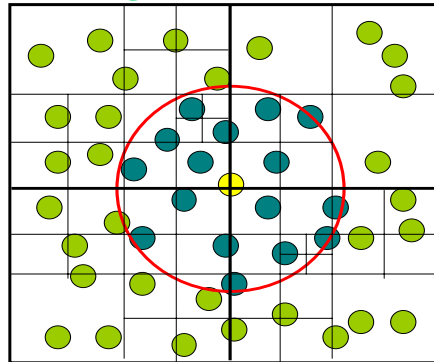
78

MolTreC ver. 1.1 (1999)

- 独自の並列ツリーコード分子動力学法
- 遠隔クーロン力もカットオフ近似しない
- 256台で約200倍高速に



Inner region = Particle-Particle
Outer region = Particle-Cell

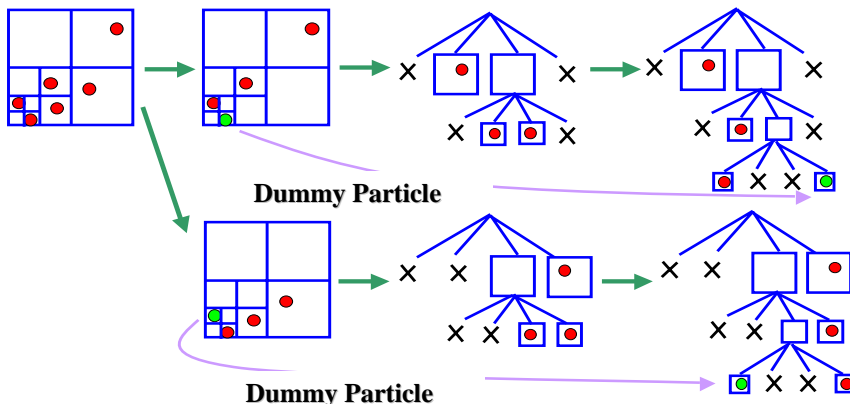


- Hierarchical tree decomposition
Barnes-Hut tree code (1986)
 $O(N^2) \rightarrow O(N \log N)$
- Outer-region: coarse updating
Saito's PPPC method (1990)

79

(HPC Asia'00)

Special Tips for Efficient Parallel "Tree construction"

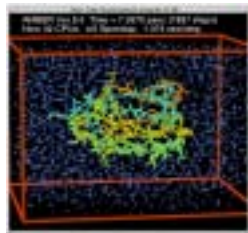


- **Tip1:** Morton key ordering
- **Tip2:** Dummy particles (to avoid dynamic memory alloc.)

80

タンパク質折れ畳み計算 MolTreC 計算時間

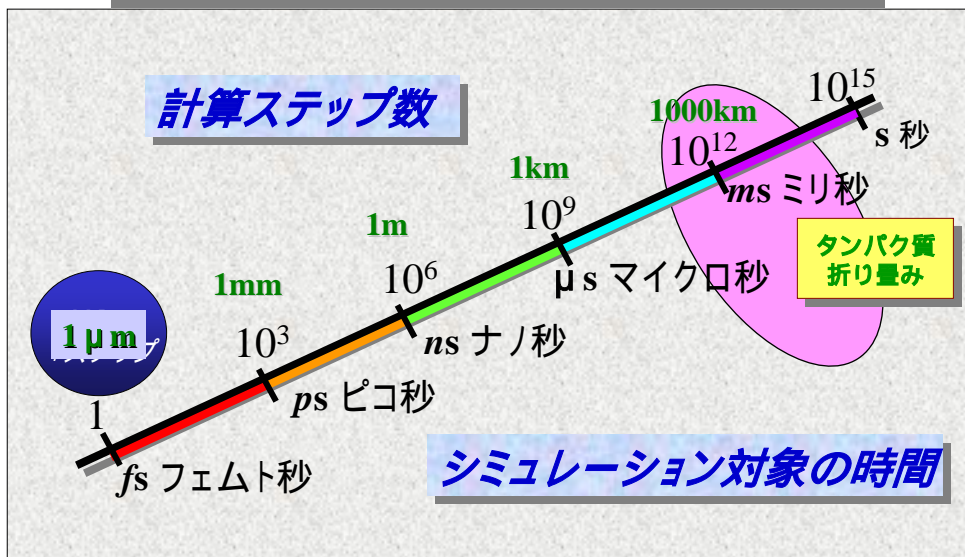
コンピュータ の台数	解きたいタンパク質の長さ	
	小型(長さ100)	標準(長さ400)
1台	10年間	160年間
250台 現状	2週間	8ヶ月間



計算時間は
知識処理による大幅
加速を併用した場合

81

分子動力学法の計算ステップ数



百万台のプロセッサで折れ畳み予測

■ IBM Research社 Deep Computing Institute

▶ “Blue Gene” プロジェクト

1GFLOPS × 100万プロセッサ = 1 Peta FLOPS
(1Peta FLOPS = 1秒間に1000兆回の演算能力)

▶ 2004年に完成予定？ BlueGene Light？

▶ タンパク質立体構造予測のオリンピック
CASP5(2002)にも出場していたが、、、、

83

MDでの並列性をどこまで出せるのか？
(プロセッサが高速になり、逆に困難に)

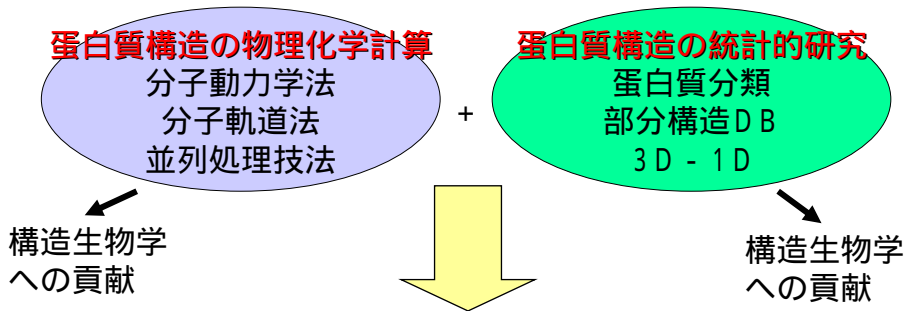
1万～5万原子	小規模(従来)
5万～10万原子	中規模(これから大事)
10万～50万原子	大規模(挑戦的だが、、)

1フェムト秒	0.1秒	中規模
1ナノ秒	28時間	160GFLOPS

- 5GFLOPS x 32 pu = 160 GFLOPS
- 160GFLOPS x 64 replica = 10 TFLOPS
- 10TFLOPS x 100条件 = 1 PFLOPS

84

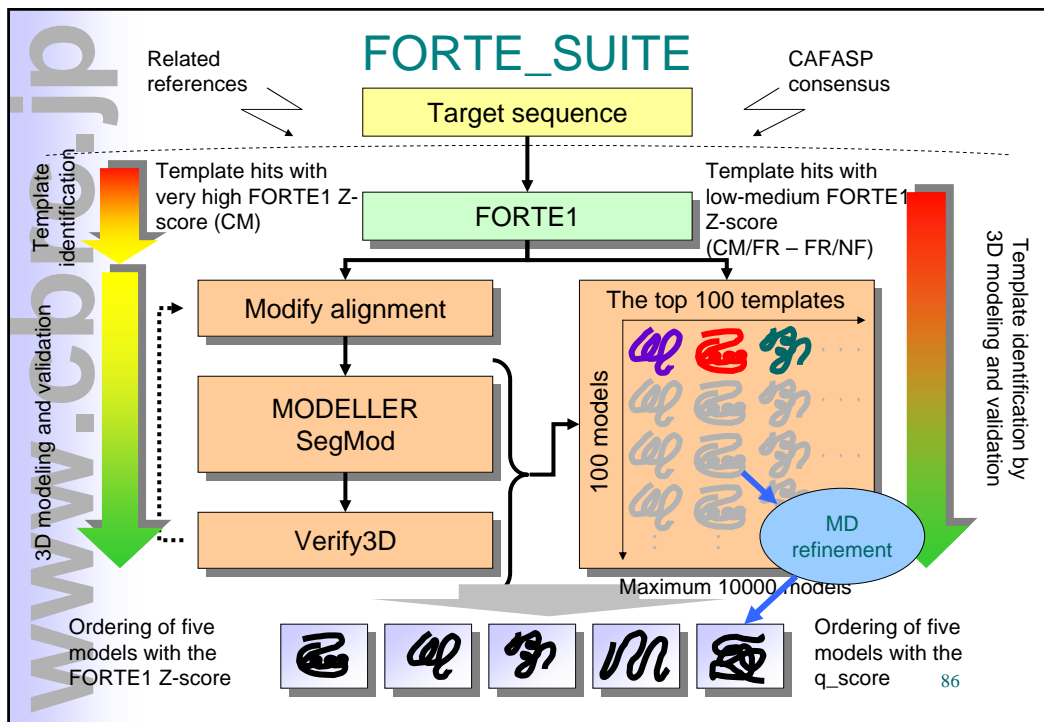
手法融合的アプローチによる構造予測



蛋白質立体構造予測法の開発
CASP5 (2002) への挑戦

日本勢としてはトップクラスでしたが...

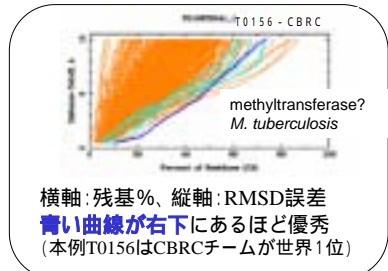
かなりかけ離れていた両分野の融合。人材と知識の合流。 85



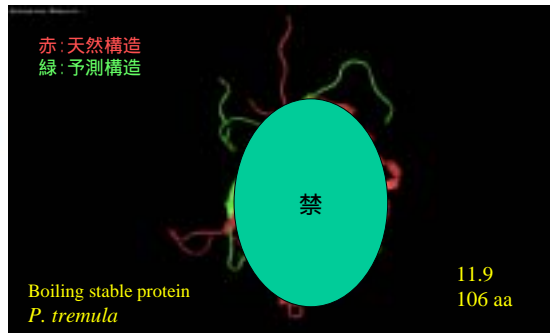
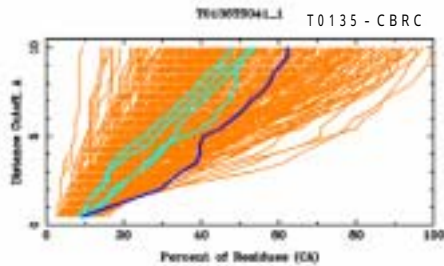
CASP5国際構造予測コンテスト(2002年夏)での客観評価

- ・2年毎開催、CASP5は216チーム参加
- ・構造が解かれる前の純粋なblind test
- ・国際的な審査員による審査

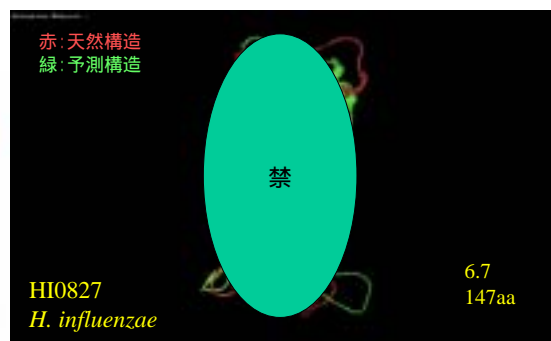
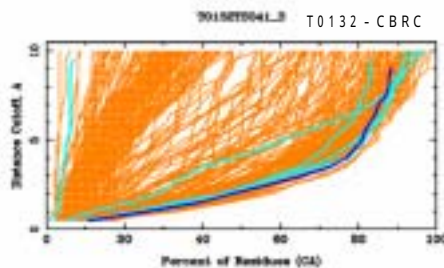
CBRCチーム及びFORTE1自動サーバ
 CASPのCM/FR部門中心に応募、国内ではトップクラス
 世界14~25位前後? (*メタサーバ群登場の影響大)
 併設のTMW部門にも参加し順調に進行中。



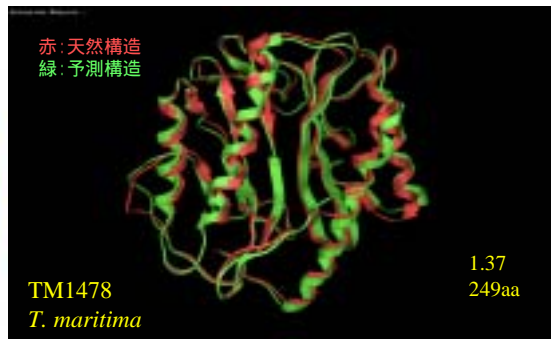
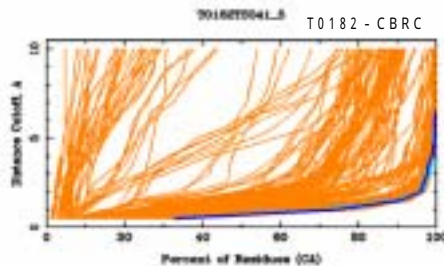
(配列相同性 ID% 10~15%)
FR級 (フォールド認識技術を要す。相同配列無し)



(配列相同性 ID% 15~30%)
CM/FR級 (配列相同性が薄い)



(配列相同性 ID% 30~50%)
CM級 (配列相同性のある立体構造がいくつか現存)





TEN MOST
WANTED

CASP番外編

<http://tmw.llnl.gov/>

Computer modellers seek out
'Ten Most Wanted' proteins
Nature **409**, 4 (4 January 2001)

- TMW0001 Rpf, (*Micrococcus luteus*, *Mycobacterium tuberculosis*)
- TMW0002 Ataxia Telangiectasia, (Human, Mouse)
- TMW0003 Telomerase Reverse Transcriptase, (Yeast)
- TMW0004 protein-lysine 6-oxidase-precursor, (Human)
- TMW0005 phosphatidylinositol kinase, (Yeast)
- TMW0006 Major Nucleocapsid N Protein, (Human respiratory syncytial virus)
- TMW0007 Ste50p, (Yeast)
- TMW0008 synuclein, alpha, (Mouse)
- TMW0009 LMP-1, (Epstein-Barr virus)
- TMW0010 DID, (*Drosophila melanogaster*)

生物学者側のリクエスト
に基づくターゲット選択

89

並列処理を利用したその他の開発例

- 並列タンパク質情報解析(PAPIA)システム
Akiyama *et al.*: *Genome Informatics*, 8, pp.131-140 (1999).
- 非冗長なタンパク質チェーン立体構造DBの自動生成PDB-REPRDB
Noguchi *et al.*: *Bioinformatics*, 16, 6, (2000).
- フラグメント近似アブイニシオ分子軌道法 ABINIT-MP
Nakano *et al.*: *Chemical Physics Letters*, 318, pp.614-618 (2000).
- 並列ペプチド配座解析プログラム ESCAPE/Hi
Ando *et al.*: *Res. Comm. in Biochem.*, 5, (1-2), pp.95-114 (2001).
- 並列質量分析プログラム
秋山ら (進行中)