PCクラスタによるゲノム解析

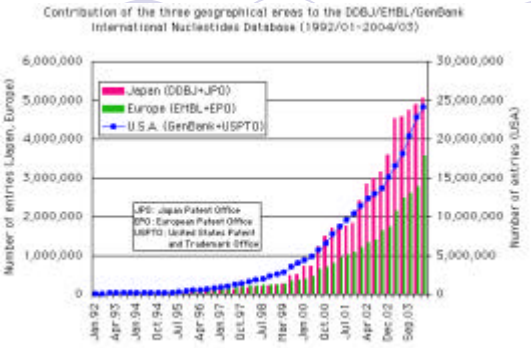
阿部　貴志
国立遺伝学研究所
生命情報・DDBJセンター
データベース運用開発研究室

Contents

- ゲノム解析の取り巻く環境
  - 日本国際塩基配列バンク（DNA DataBank of Japan ;DDBJ）
  - ゲノムプロジェクトの現状
- ゲノム解析例
  - 情報資源の信頼性 の向上にむけて
    (比較ゲノムプロジェクト：CGM)
    - 100種類以上の微生物ゲノムに対して統一的なアノテーションを実施してゲノムデータを評価
  - 生物多様性の解明にむけて
    - 自己組織化地図法(Self-Organizing Map ;SOM)によるゲノム解析
      - ゲノム配列に潜む生物種固有の特徴抽出

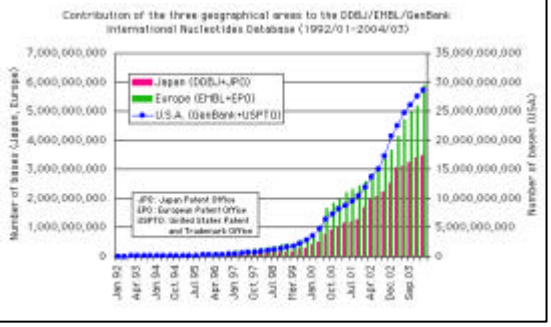
ゲノム解析の取り巻く環境

- 日本国際塩基配列バンク
  - DNA DataBank of Japan（DDBJ）
- ゲノムプロジェクトの現状



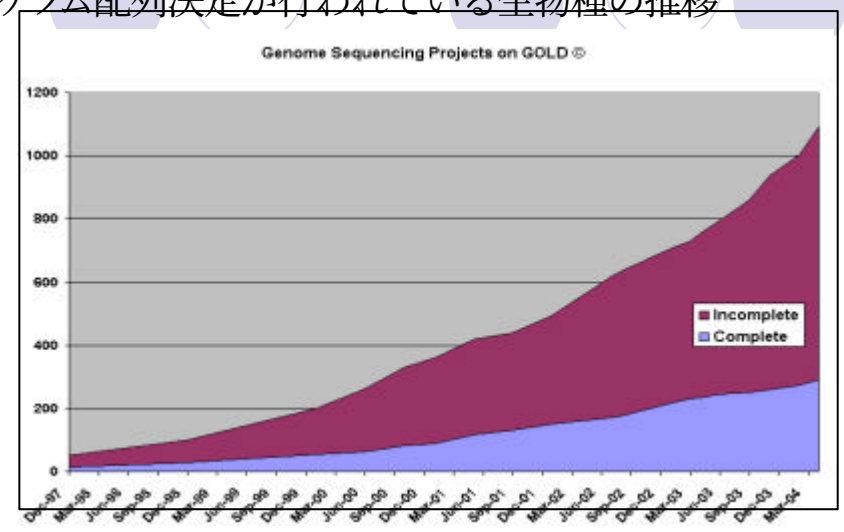
DNA塩基配列データの
登録数と登録塩基数の推移

DDBJデータベースの拡大
2000年の末に1000万件100億塩基対に達した塩基配列データベースの規模は、2003年に3000万件360億塩基対を超えた。今後も、毎年少なくとも50％程度は伸び続けていくものと予想される。



ゲノムプロジェクトの現状

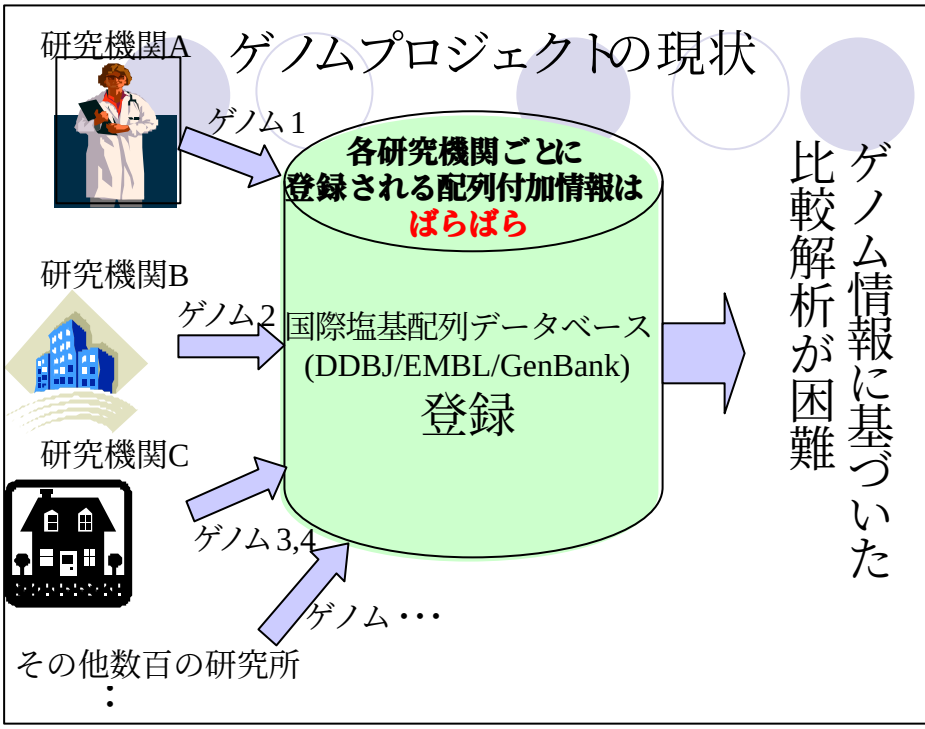
ゲノム配列決定が行われている生物種の推移



現在、1,000種を越える生物種のゲノム配列の解読が行われている。

Genome OnLine Database (http://www.genomesonline.org/)より

ゲノムプロジェクトの現状



国際塩基配列から見たPCクラスタやグリッドへの期待

- 3000万件のレコード（1.5倍／年）の査定（人の判断が必要）
- バックアップ（公開／未公開）
- サービスのためのミラーリング
- 生物情報の統合的な研究リソースの作成と提供

情報資源の信頼性 の向上にむけて (比較ゲノムプロジェクト:CGM)

- 微生物の完全長ゲノムにつけられている配列付加情報の網羅的な見なおし
- 統一的な配列付加情報(アノテーション)の付与
- 標準ORFを決定するための基本手順の開発
- 「第3者アノテーション」を支援するためのソフトウェアの開発

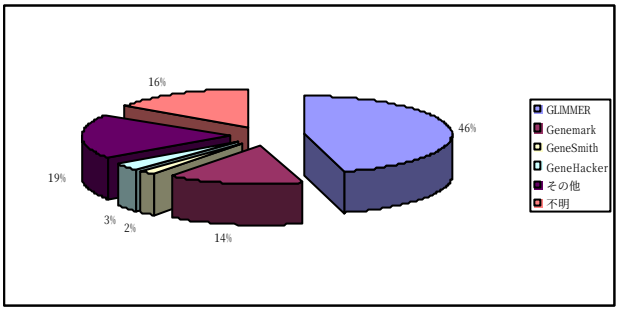
Why retrospective evaluation ?

Protocols for annotation were diverse:

- ORF prediction
  - Diversity of prediction programs
  - Diversity of parameters, e.g. the minimum length of ORFs
- Reference databases
  - Genes in DDBJ/EMBL/GenBank are continuously expanded and updated
  - Reference databases like InterPro are also continuously expanded and updated

Diversity of prediction programs and a parameter

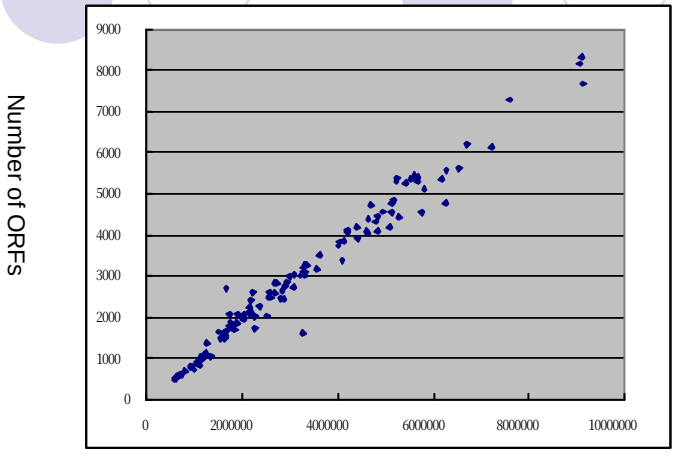
ORF Prediction programs used



Threshold value to determine the ORF

>5aa	1
>20aa	1
>30aa	24
>33aa	1
>33.3aa(100bp)	3
>40aa	1
>50aa	8
>60aa	4
>66.6aa	1
>80aa	2
>100aa	6
>150aa	1
>200aa	1
>300aa	1
>400aa	1

Number of ORFs vs genome size



Microbial genomes in INSD*

(*) The International Nucleotide Sequence Databases (DDBJ/EMBL/GenBank)

Archaea	19
Bacteria	157
Eukaryota	6
Total	182

(as of July 27th, 2004)

http://gib.genes.nig.ac.jp

Description of an ORF (CDS)

Species	Name	Start	End	Category
Ecol_K12_MG1655:	b1432	1501681	1502889	Hypothetical proteins

CDS	
codon_start	1
db_xref	GI:1787702
function	putative factor; Not classified
gene	b1432
location	1501681..1502889
note	o402; This 402 aa ORF is 30 pct identical (9 gaps) to 105 residues of an approx. 120 aa protein VSDF_SALDU SW: P24421
product	
protein_id	
transl_table	
sequence	
links	

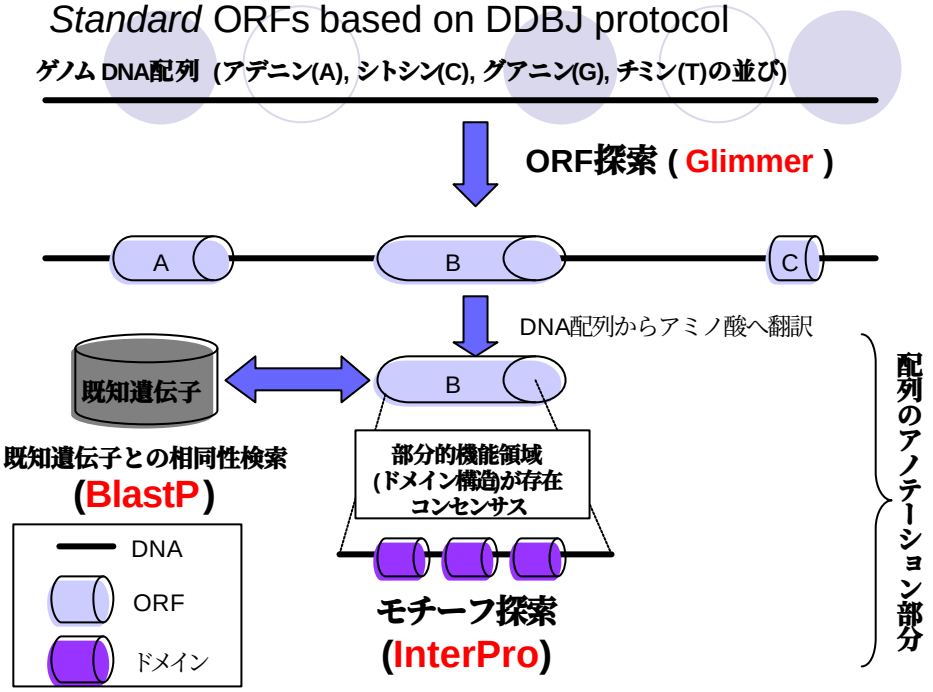
Agrobacterium tumefaciens C58 (Cereon)

Feature Information

Species	Name	Start	End	Category
Atum_C58_CERBON	AGR_C_884	489349	489630	Hypothetical proteins

CDS	
codon_start	1
db_xref	GI:15155429
gene	AGR_C_884
location	489349..489630
note	hypothetical protein
product	AGR_C_884p
protein_id	AAK86313.1
transl_table	11
Links	OTOP (Genome TO Protein structure and function) ORORT

Standard ORFs based on DDBJ protocol

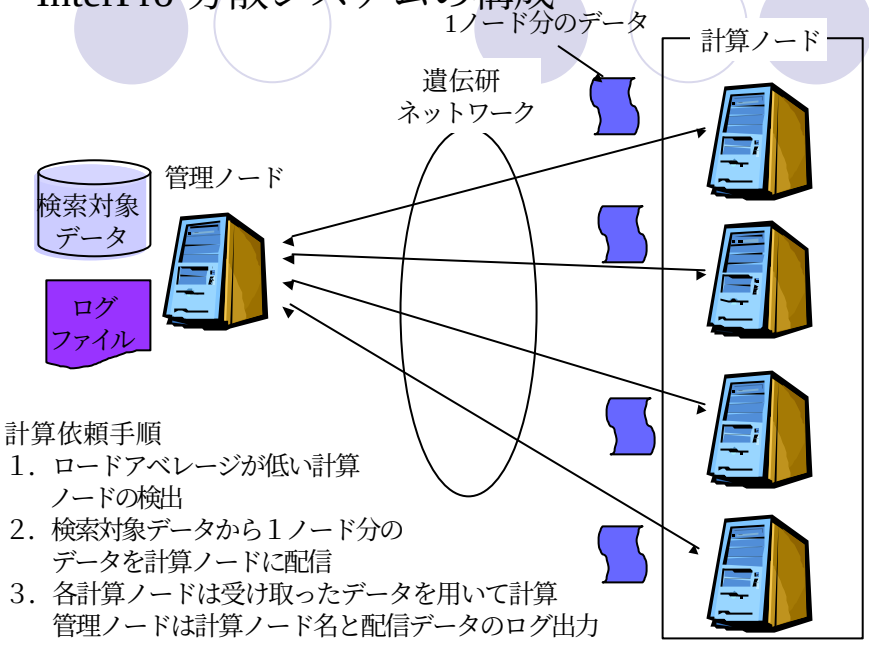


Standard ORFs based on DDBJ protocol に使用したプログラム一覧

	プログラム名	プログラムの機能
1	Glimmer	遺伝子領域探索プログラム
2	BlastP	アミノ酸配列に基づく相同性領域の探索
3	InterPro	遺伝子領域内のモチーフ探索

この protocaol では、InterProの計算時間が他と比べ、圧倒的にかかる。そのため、PCクラスタ上での分散環境を構築し、実行を行っている。

InterPro 分散システムの構成



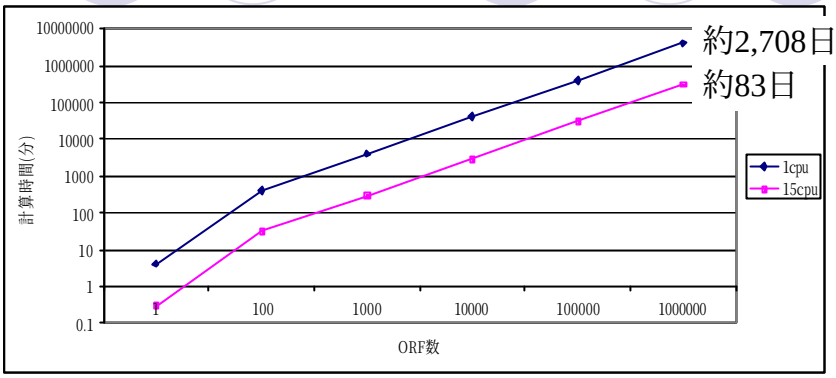
計算依頼手順
1．ロードアベレージが低い計算ノードの検出
2．検索対象データから１ノード分のデータを計算ノードに配信
3．各計算ノードは受け取ったデータを用いて計算
管理ノードは計算ノード名と配信データのログ出力

InterPro の実行時間

- テスト環境　Pentium4 2.2GHz * 15CPU
- アプリケーション　InterProScan
- 検索対象データ　大腸菌の予測ORF 1584

- 実行結果
  - 15並列実行　約 8時間
  - 1ノードで実行　約 105時間
  - 並列度　約 13倍

InterPro の計算時間とデータ件数との関係



現状では、約150万件程のORFが対象。
今後、年1.5倍ほどの増加が予想されている。
⇒より強大なマシンリソースが必要(Grid, etc)

生物多様性の解明にむけて

- 自己組織化地図法によるゲノム解析
  (Self-Organizing Map : SOM)
  - ゲノム配列に潜む生物種固有の特徴抽出

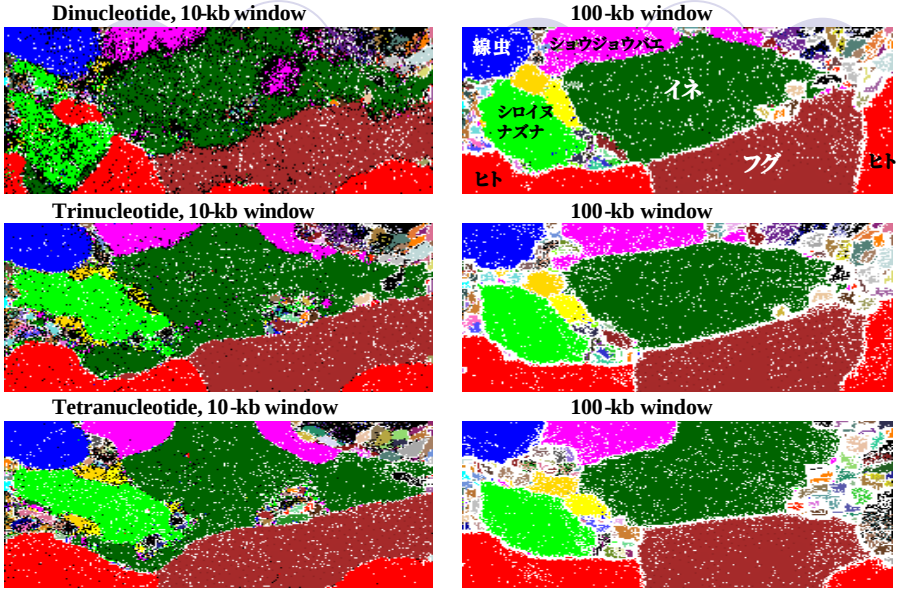
ゲノム解析における重要な課題

大量なゲノム配列からいかに多くの知識を効率的に発掘・発見してゆくか？

大量なゲノム情報の全体像と部分情報の両方を効率的に把握するための情報学的な手法の開発が必要

自己組織化マップ
(Self-Organizing Map : SOM)

ゲノム情報の解析に用いるために、入力データの順序に依存しない形に特化させた



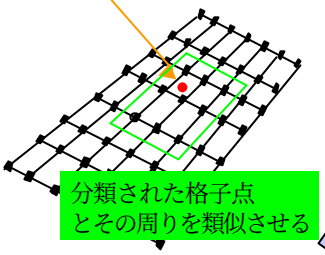
原核生物81種と真核生物9種を用いたときのSOM解析

生物種の情報を与えることなく連続塩基の特徴のみで分類が可能

通常のSOM

大統領の原稿10,000件
●, ▲, □, △, ■, …..

一番近いところを探す



分類された格子点
とその周りを類似させる

30 x 30
2次元上の格子

SOM for genome informatics

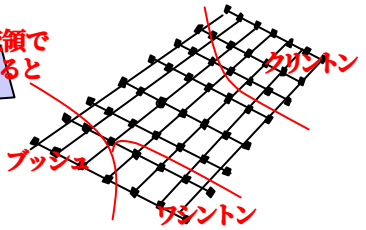
Initial vectors were determined with PCA.
Batch-learning SOM was conducted.

例えば20人の大統領の各500件で、合計10,000件の演説原稿について、1000種類の単語 (Science, War, Iraq, Peace, ..) に着目したSOMを考える。

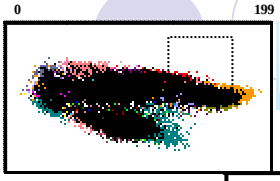
初期状態としては、大統領とは無関係の900件の原稿を集めて、各原稿を任意の順番で、30 x 30の2次元の格子上に並べた上で、1000種類の単語の使用頻度を計算する。

α(t)　例えば0.95 ; β(t)　例えば５

学習後、大統領で結果をみてみると

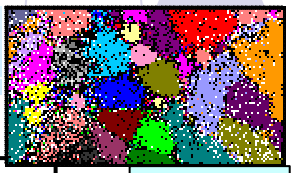


ゲノム情報の解析に用いるために、入力データの順序に依存しない形に特化させたアルゴリズム

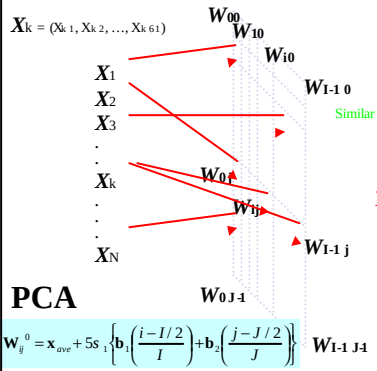


BLSOM
(Batch-Learning SOM)

Iteration (t = 1, 2,…,T)

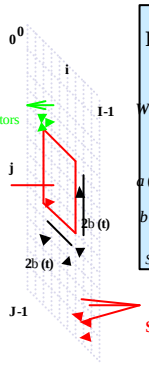


Initializing Weight Vectors → Classification of Input Vectors → Updating Weight Vectors → Mapping of Genes



PCA

$$W_{ij}^{0} = x_{ave} + 5s_1\left\{b_1\left(\frac{i-I/2}{I}\right) + b_2\left(\frac{j-J/2}{J}\right)\right\}$$



Parameters in learning process

$$W_{ij}^{(new)} = W_{ij}^{(old)} + a(t)\times\frac{\sum_{x_k \in S_{ij}} x_k}{N_{ij}} - W_{ij}^{(old)}$$

$$a(t) = \max\left\{0.01, a_{init}\left(1-\frac{t}{T}\right)\right\}$$

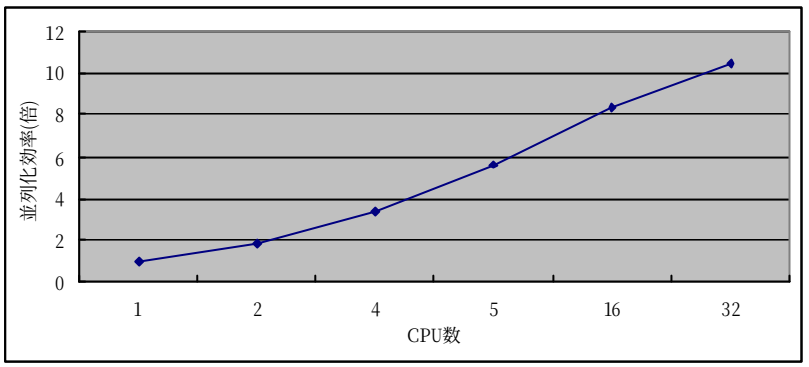
$$b(t) = \max\{0, b_{init} - t\}$$

S_{ij}={Wij | i-β(t)<i'<i+β(t), j-β(t)<j'<j+β(t) }

並列化が可能

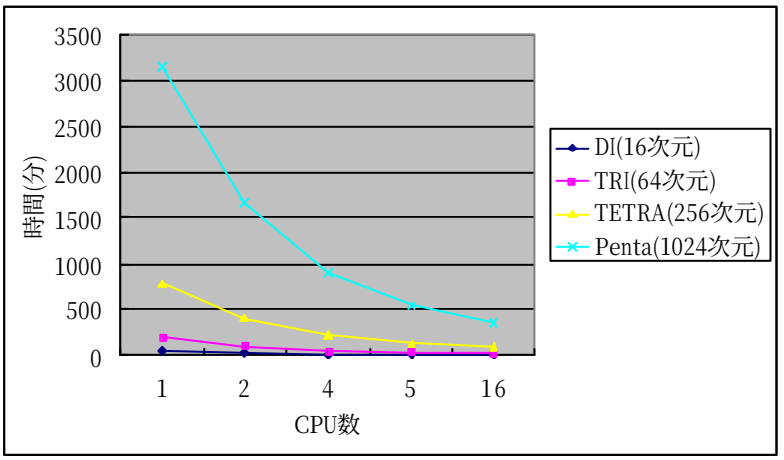
自己組織化地図法の並列化効率

- テスト環境 :ITBL 計算機 (CPU: 675MHz)
  - データセット数： 297,873, 次元数 :16次元 (2連続塩基)
  - 計算時間は、データセット数と次元の増加数に比例



並列化プログラミングはMPIを使用。

自己組織化地図法の次元数と計算時間の関係(1サイクル時)



通常、100サイクル程度繰り返し、学習を行う。

ゲノム配列が解読された約170種の生物種の 2～5連続塩基頻度のSOMを作成しており、予想を遥かに超える分離能を得ている。5連続塩基の場合、国立遺伝学研究所のスーパーコンピュータの32 CPUを用いた並列計算でも、10日程度の長大な計算が必要になる。

重要な生物学的特長の検出 (シグナル配列の網羅的な探索) のためには、さらに大規模な計算が避けられない。

環境微生物を中心とした難培養性微生物の、計算機上での系統分類と、新規な有用ゲノムの探索を行う。

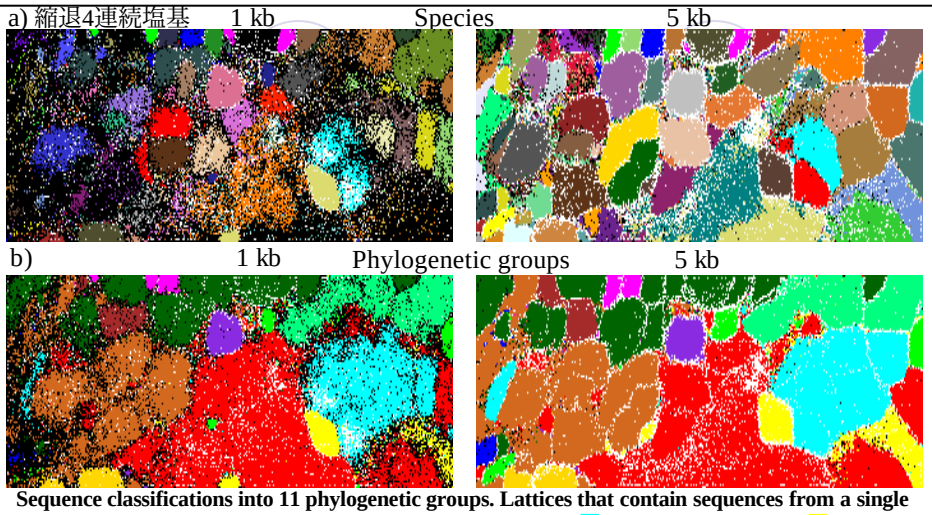
応用例： Metagenome (膨大な未開拓ゲノム資源の活用)

難培養性微生物 、もしくは共生 •寄生生物系のDNAを自然環境から微生物の培養 •分離をすることなく、直接回収し、配列決定を行い、産業的に有用な遺伝子を探索する手法

Metagenome 問題点

- **DNA配列の由来を推定することは困難**
- **多種多様な環境中での共生細菌等の多様性の解明が困難**

- SOMを系統分類法に適用し、新規性の高い未知微生物種を効率的に推定する方法を開発した。



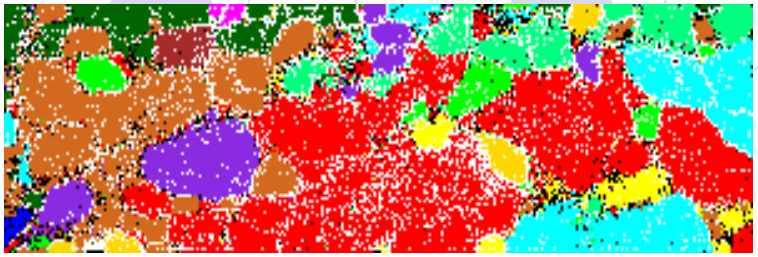
Sequence classifications into 11 phylogenetic groups. Lattices that contain sequences from a single group are indicated in color as follows: Alphaproteobacteria (■), Betaproteobacteria (■), Deltaproteobacteria (■), Gammaproteobacteria (■), Archaea (■), Chlamydia (■), Firmicutes (■), Actinobacteria (■), Fusobacteria (■), Thermotogae (■), Other(■).

DegeTetra-SOMs for 1-kb and 5-kb sequences from 81 bacterial genomes

縮退では、ゲノム配列の相補性の影響を除去するために、相補的な配列 (例えば、AAとTT) を同一のものとみなし、頻度解析を行う。

データセット数 : 168,680件, 次元数: 136次元, 計算時間(16cpu時): 577.6m

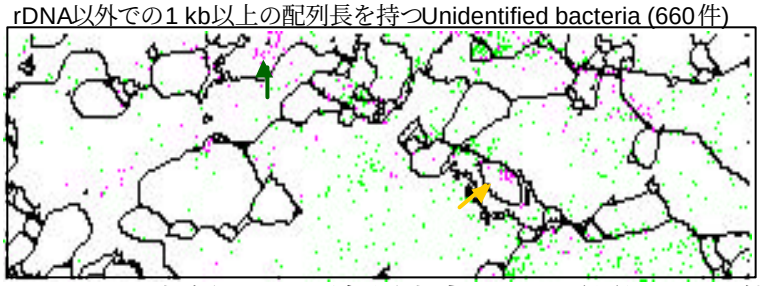
配列既知な原核生物147種での断片化サイズ5-kb, 縮退4連続でのSOM



コンプリート 108種
ドラフト 39種

Alphaproteobacteria (■), Betaproteobacteria (■), Deltaproteobacteria (■), Gammaproteobacteria (■), Archaea (■), Chlamydia (■), Firmicutes (■), Actinobacteria (■), Fusobacteria (■), Thermotogae (■), Other(■).

rDNA以外での1 kb以上の配列長を持つUnidentified bacteria (660件)



GenBankに収録された反芻胃由来のDNA配列 (■ 329件)

■：メタン生成菌、■：Desulfovibrio desulfuricans硫酸塩還元細菌

共に嫌気性菌であり、反芻胃の中にいると考えて整合性のある菌である。

自己組織化マップを系統推定法へ適応

- DNA配列の系統群の推定が可能
- 多種多様な環境中の共生細菌等の多様性の可視化が可能

深海の火山口付近のような極限環境を含む様々な環境に生息する微生物の多様性の解明が可能

生体内の腸内細菌の多様性や感染症を引き起こす微生物の識別などにも役立たせることが可能

Collaborators

- **比較ゲノムプロジェクト: CGM**

Bioinformatics Group, RIKEN GSC
Akihiko KONAGAYA
Fumikazu KONISHI
Shingo OKI
Hiroshi UMEDA (IBM)

Graduate School of Knowledge Science, Japan Advanced Institute of Science and Technology (JAIST)
Kenji SATOU
Shinichi TSUJI (NEC SOFTWARE HOKURIKU)
Yasuhiko NAKASHIMA (NEC SOFTWARE HOKURIKU)

Tokyo Medical and Dental University
Toshinori ENDO (Hokkaido University)
Tekehiro FURUDATE (Hitachi SK)

Japan Science and Technology Agency
Masako KURODA
Kyoko SUZUKI
Toshiyuki KOIKE
Shinya NODAGUCHI
Munakata YOSHIHISA
Shunji KOHNO (Hitachi)
Shinsuke DOHKAN (Hitachi)

Center for Information Biology and DDBJ, National Institute of Genetics (NIG)
Hideaki SUGAWARA
Satoru MIYAZAKI (Tokyo Science University)
Masahito YAMAGUCHI (Fujitsu)
Yasumasa SHIGEMOTO (Fujistu)
Masashi MATSUO (Fujitsu)
Keiichi IDA (Fujitsu ISD)
Kazutaka SUGIMOTO (Fujistu ISD)

- **自己組織化地図法(Self-Organizing Map：SOM)によるゲノム解析**

The Graduate University of Advanced Studies
Toshimichi IKEMURA
Nara Institute of Science and Technology
Shigehiko KANAYA

Yamagata University
Makoto KINOUCHI
National Institute of Genetics (NIG)
Hideaki SUGAWARA
Takashi ABE

Acknowledgements

- This work has been partly supported by BIRD of Japan Science and Technology Agency (JST) and also partly by the Grant-in-Aid for Scientific Research on Priority Area "Genome Information Science", Ministry of Education, Sports, and Science (MEXT), Japan