並列推論マシンを用いた
遺伝子情報処理

ICOT
第７研究室

新　田　克　己

概　要

並列推論マシンの応用としての遺伝子情報処理の研究についての報告

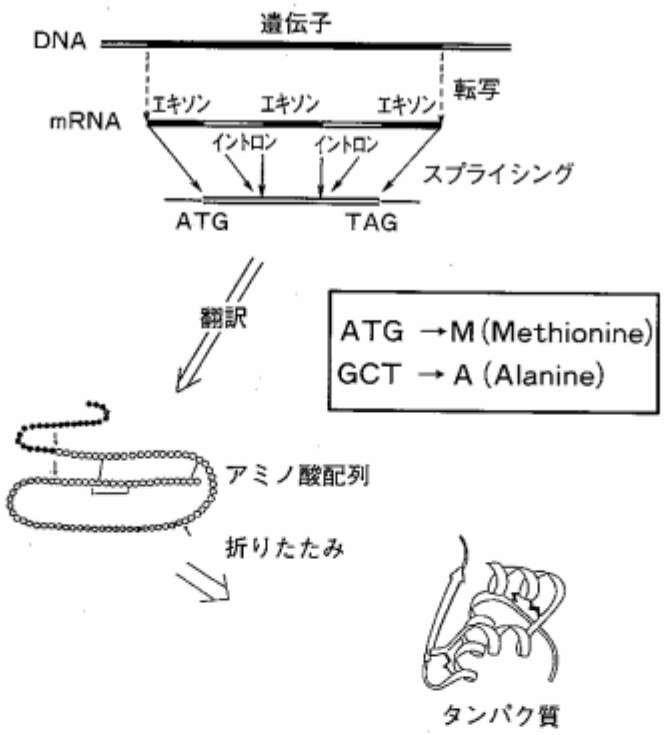
主な内容

１．遺伝子とタンパク質の基礎知識
- 遺伝子とタンパク質
- タンパク質の立体構造
- 核酸塩基，アミノ酸配列データベース

２．遺伝子情報処理の重要性と５Ｇ技術との適合性

３．遺伝子情報処理の研究計画
- 統合的な分子生物データベース
- タンパク質の構造予測・機能予測

遺伝子とタンパク質



タンパク質の構造

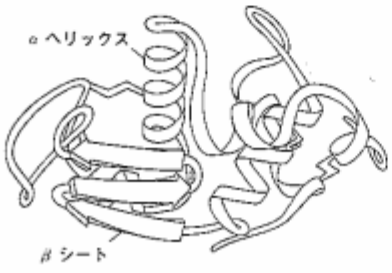
1 次構造

例　Val　Leu　Ala　Ser　Gly　Arg　Ala　Val ····

2 次構造

例　α ヘリックス, β シート, ターン, コイル

立体構造



構造間の関係付けは，
自然言語理解における
表層文 ⟷ 意味　の関
係付けと類似。

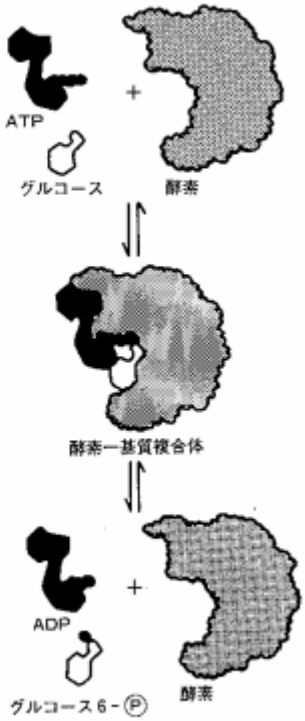
⇨ 知識処理の手法が有効

タンパク質の立体構造と機能

タンパク質の立体構造と機能の間には密接な関係がある。

例　グルコースとＡＴＰが酵素の活性部位に結合して反応し，ＡＤＰとグルコース６－Ⓟを生成。

(ワトソン「遺伝子の分子生物学」《トッパン》による)



DNA・タンパク質データベース

DNAデータベースの例

GenBank	(核酸の１次構造)	Los Alamos 国立研究所
EMBL	(核酸の１次構造)	欧州分子生物学研究所
DDBJ	(核酸の１次構造)	国立遺伝学研究所

タンパク質データベースの例

PIR (タンパク質の１次構造)	米国生物医学研究財団
PDB (タンパク質の立体構造)	Brookhaven 国立研究所

データベースの特徴

実験データ／報告の集積 ⇨ 分類，体系化が大きな問題

構造 ⇔ 機能の対比によるデータの意味付けが重要

遺伝子配列データベースGenBank

階層的なデータ構造

属性１つに複数データ（著者，文献など）

→ 従来型の関係ＤＢへの格納には不向き

急速な膨張と改良

データ構成に変更が多い

→ 従来型の関係ＤＢでは変更への対応が困難

⇨ ICOTの非正規型関係DB（Kappa）が適合

```
REFERENCE   1  (D
  AUTHORS   Goodwin,E.B.,
  TITLE     Cloning and chara
            regulatory myosin li
  JOURNAL   J. Biol. Chem. 262, 110
  STANDARD  simple staff_review
FEATURES             from  to/span     descri
     pept              82       555     regulat
     refnumbr           1         1     numberec
BASE COUNT         237 a    155 c    166 g
ORIGIN      EcoRI site.
        1 aattcctagt ctacagctgt gcacggtcg
       61 atcaactgaa gaggagtcaa aatggccga
      121 ccccagaaac aaattcagga gatgaaag
      181 ggtttcgtca gcaaagagga tatcaaa
      241 gacaaagaat tgacagctat gttga
      301 ttgtccatct tctcagacaa gct
      361 ttcgccatgt ttgatgaaca g
      421 ttagaaaaca tgggaga
      481 cctgtggaag ga
      541 gagga
```

遺伝子情報処理の重要性と5G技術との適合性

応用の重要性

- タンパク質の設計　　例　インシュリン
- 遺伝病・癌の解明　　例　鎌型赤血球貧血病の解明
- 生命の起源の解明

情報量の爆発

- 年々，増加するデータ（多種，多量の生データ）
  50万base（1982） ⇨ 500万base（1985） ⇨ 5000万base（1990）

５Ｇ技術（並列知識情報処理）との適合性

- 機能予測，構造予測 → 構造と機能の意味処理的マッチング（並列推論）
- 多量の生データ → データの分類，体系化（知識ベース化）
- 大量の計算処理 → 記号処理と数値処理の混在（並列ソフトウェア）

5G技術に基く遺伝子情報処理の研究内容

1．統合的な分子生物データベース構築を目指す研究

現状：多くのデータベースにモデルがない

研究項目

- 遺伝子データベースのモデリング ⇨ 論理に基づくデータモデルが有望
- 種々の実体（系統発生，種，遺伝子，RNA，タンパク質，分子等）を含む知識ベース／データベースの研究

研究計画

- Kappa上に統合的な知識ベース／データベースを試作

  応　　用 ……分子生物データベース

  知識ベース ……知識表現言語

  データベース……並列データベース管理システム

平成元年度の成果

- Kappa上にGenBankを実装

  GenBankの約１／４（7,285遺伝子＋付加情報）

  → 3,000万ワード分の記憶領域に格納

  （インデクスは10個付加）

  PSI-Ⅱの２次記憶に格納

  ４つのテーブルに分けて格納

  （従来の関係DBでは38テーブル必要）

2．タンパク質の機能予測・構造予測を目指す研究

現状：一次構造がわかっていても，高次構造や機能のわかっているタンパク質は全体の数パーセントにすぎない。

研究項目

一次構造からのタンパク質の高次構造予測・機能予測。

- 意味のある配列パターンは何か
- 配列パターンと構造・機能との関係は何か

研究計画

① タンパク質を進化，機能，構造などの観点より分類し，保存パターンを抽出 ⇨ 並列処理による高速化が必要

② 保存パターンと構造・機能との関係を抽出

⇨ 構造・機能に関する制約条件を利用

確率的な手法

仮説の生成と検証

- 応　　用……タンパク質の構造・機能予測
- 知識処理技術……知識表現，推論
- 並列処理技術……KL1プログラミング技法

平成元年度の成果

- 相同 sequenceの検索

  DPマッチングアルゴリズムをKL1とA'UMで記述

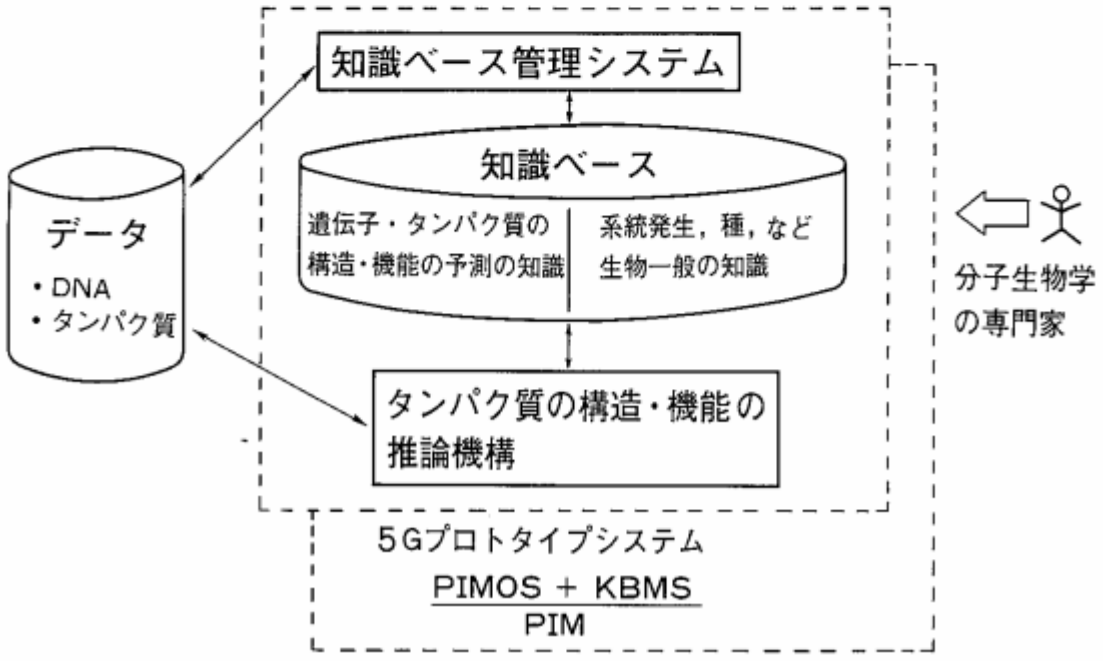
- 保存パターン(モチーフ)によるタンパク質の分類

  モチーフの例

  NGFGRIGR (130-134) SNASCTTNCLAP (12) EGLMITVH (2) TATQKTVD (20) STGAAKAVGKV (5) GKLTGMAFR (75-77) SWYDNE

  アミノ酸配列が特定のモチーフを含んでいるか否かにより タンパク質を分類

遺伝子の研究支援システムの構成例



平成２年度の計画

- 統合的データベースの研究
  - ― 既存データベースの技術調査と代謝反応データベース等の記述実験
  - ― GenBankとPIRを含む「統合データベース」の試作

- 構造予測のための研究
  - ― モチーフの抽出手法の検討
  - ― 実験・評価用ソフトウェアツールの開発
    （京大，化学研究所との共同研究）

- 米国アルゴンヌ国立研究所との共同研究の実施